Discovering pathways to autism spectrum disorder by using functional and integrative genomics approaches to assess monozygotic twin differences

Ayden Saffari

Department of Psychological Sciences,
Birkbeck, University of London

A thesis submitted for the degree of
Doctor of Philosophy
March 2017

This thesis describes research carried out by Ayden Saffari over the course of his PhD studentship at Birkbeck, University of London. The work is the author's own, except for the following collaborations and contributions:

Overall:

The cohort of twins used in this study, and the associated biological samples and behavioral data utilized, are from the Twins Early Development Study (TEDS), a longitudinal study investigating the cognitive and behavioural development of twins born in the UK between January 1994 and December 1996 (cited in main text). TEDS is a project of the MRC Social, Genetic and Developmental Psychiatry Centre, Institute of of Psychiatry, Psychology and Psychology and Neuroscience, King's College London, and is led by Professor Robert Plomin.

Chapter 2: Gene expression profiling of monozygotic twins with autism spectrum disorder

Professor Robert Plomin[1] (principal investigator - TEDS) –Study design (for the original longitudinal study)

Dr. Emma L. Meaburn[2,1] - Study design, sample collection, Human Gene expression array experiments, data generation

Dr. Angelica Ronald[2,1] - Subject selection

Chapter 3: Characterising the transcriptome in autism spectrum disorder using RNA-seq

Dr. Emma L. Meaburn[2,1] (principal investigator) - Study design, sample collection

Dr. Matt Arno[3] (head - Genomics Centre) - RNA isolation, RNA-seq library preparation

Dr. Alka Saxena[4] (head - BRC Genomics Core Facility) – HiSeq sequencing (pilot study), data deconvolution

Illumina, UK Ltd –HiSeq sequencing (full study)

Chapter 4: Estimation of a significance threshold for Methylation-Wide Association Studies

Professor Frank Dudbridge[5] (principal investigator) - Study design

A number of public 450K methylation datasets also used – see main text.

Chapter 5: Integrating multi-dimensional omics datasets

Dr. Chloe C.Y. Wong[1] – 27K Methylation data generation *(N.B. this data is from a published methylation study on the same sample of twins which was made available for this study –see main text)*

1 King's College London, MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, De Crespigny Park, London, United Kingdom
2 Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom
3 Genomics Centre, School of Biomedical and Health Sciences, King's College London, London, United Kingdom
4 BRC Genomics Core Facility, King's College London, Guy's Hospital, London, United Kingdom
5 Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Signature of supervisor [redacted]

Abstract

Autism spectrum disorder (ASD) is a common developmental disorder typified by deficits in social communication and stereotyped behaviours. Despite evidence of a strong genetic basis to the disorder, molecular studies have thus far had little success in identifying risk variants or other biomarkers, and presently there is no unified pathomechanistic explanation. Monozygotic (MZ) twins show incomplete concordance in autistic traits, which suggests that alternative risk pathways involving non-shared environmental (NSE) factors could also have an important role to play in ASD. In this thesis, we describe microarray and RNA-seq studies characterising gene expression in a sample of 53 ASD MZ twin pairs from TEDS. The overall aims were to: 1) establish convergent evidence for genes and pathways involved in the etiology of ASD comparing affected and unaffected subjects across the sample 2) to identify those responsive to the environment by examining differences within the discordant pairs. We found a number of genes were differentially expressed including *DEPDC1B* - the most significant finding in cases vs controls, which also showed consistent down regulation within pairs. We further identified *IGHG4, IGHG3, IGHV3-66*, *HSPA8P14, HSPA13*, *SLC15A2*, and found that these results were enriched for transcriptional control, immune, and *PI3K/AKT* signalling pathways. We suggest that as these were found to be perturbed in the discordant twins, they could represent ASD risk pathways sensitive to the NSE. Next, we investigated integrative genomics methods for performing meta-dimensional analysis using the expression data along with methylation data on the same cohort. After applying regression-based joint analysis methods, and meta-analysis p-value combination methods to our datasets, a number of genes obtained nominal significance across the datasets, including potential genes of interest: *NLGN2*, *UBE3A*, *OXTR*. We suggest these represent genes with evidence for being functionally relevant to ASD.

Acknowledgements

First and foremost, I'd like to express my sincerest gratitude to my supervisors Emma and Frank for their guidance, support and enthusiasm throughout the project, which has been essential to its completion. I would also like to thank the following people: Prof. Robert Plomin for kindly allowing the TEDS data to be used for this project, and of course the TEDS families for their continuing participation in the study, Dr. Matt J. Silver (LSHTM - MRC International Nutrition Group) for his support and for proofreading some of the material contained within, Dr. Vincent Plagnol (UCL - Genetics Institute), Prof. Michael A. Simpson and Dr. Nikolaos Barkas (KCL - Division of Genetics and Molecular Medicine) for their advice on methodological aspects of the RNA-seq study. And finally, Dr. Denise McDonald (QUB - Centre for Experimental Medicine) and Dr. Kostas A. Papageorgiou (QUB - School of Psychology) for their advice on academic matters and thesis writing and general all round encouragement. I am also grateful to the following for funding this research: Bloomsbury Colleges (PhD studentship award), The Royal Society (Research grant - RNA-seq experiment).

For my parents, sister, close family and friends, thank you for your love and support.

Contents

1 Introduction

1.1 Autism spectrum disorder

Leo Kanner first noted "innate disturbance of affective contact" in 11 children [1], which he termed "early infantile autism" [2]. While Kanner's initial description still captures some of the essential features of autism, the precise clinical definition has been continually revised and updated to reflect advances in our understanding. The latest edition of the diagnostic and statistical manual of mental disorders V (DSM-V) classifies autism spectrum disorder (ASD) as a neurodevelopmental condition featuring two core symptoms, 1) deficits in social communication and 2) stereotyped behaviour [3]. This supersedes the previous DSM-IV definition which used as its main diagnostic criteria what came to be known as the "classic triad" of 1) impaired reciprocal social interaction, 2) deficits in language and communication development, and 3) restricted repetitive behaviours and interests [4].

Perhaps one of the defining characteristics of ASD is its phenotypic heterogeneity, with individuals varying in the severity and extent of impairments in language ability and cognitive development. ASD is often comorbid with other conditions such as intellectual disability (in over 70% of ASD affected individuals) [5], macrocephaly (in around 20% of ASD cases) [6], epilepsy (~ 25%) [7], as well as a number of rarer conditions. The DSM-IV previously used ASD as an umbrella term bringing together a number of clinically distinct neurodevelopmental disorders including autism, Asperger's and pervasive developmental disorder not otherwise defined (PDD-NOS). The DSM-V now uses ASD as a singular classification, with

provisions for the inclusion of patient-specific characteristics outside of the core symptoms, for example, the presence of a genetic syndrome with known autism comorbidity [8]. This is believed to more accurately reflect the spectrum nature of autism, as well as the underlying biology [9, 8].

Prevalence of ASD has rapidly risen over the past several decades, from a rate of 1 in 10,000 in the 1960's [10], rising to 1 in 2500 in 1980's (at time of the introduction of the DSM-III) [11], with current estimates between 1 in 150 [12] to 1 in 100 [5, 13]. There is also a well-established gender imbalance, with males being 4 times as likely as females to be affected [5]. There is some debate as to whether the increase in incidence is due to the changing and broadening of diagnostic criteria, screening efficacy, increased awareness, or increased exposure to environmental risks factors [14] – some of which could be linked to changes in modern lifestyles such as increasing paternal and maternal age [15]. But before we can begin to pick apart the etiology, we must first form a better understanding of the underlying pathomechanisms. This is perhaps where research into ASD lags behind other neurodevelopmental and psychiatric disorders, as currently there is no coherent narrative running from etiology and disrupted cellular processes, through to developmental disruption and final manifestation of autistic behaviours.

1.2 Insights from genetics

1.2.1 Quantitative studies

ASD has been described as the "most genetic" neuropsychiatric or developmental disorder, based on accumulated findings from several decade's worth of quantitative research. Family studies have shown sibling relative risk to be between 20 and 50 times greater than in the general population [16, 17], and subclinical autistic traits such as deficits in social cognition and language are also commonly observed at higher rates in first-degree relatives [18, 19](sometimes referred to as the "broader phenotype").

Twin studies estimate the proportion of liability in ASD attributable to genetic factors (the narrow sense heritability) by comparing concordance rates between monozygotic (MZ) and dizygotic (DZ) twins. Some early studies estimated this at between 60 and 90% [20, 21, 22] based on concordance rates of 36% in MZs compared to 0% in DZs [20], 91% to 0% [21], and 60% to 0% [22]. Two more recent studies revisited twins-based heritability estimation using more up-to-date diagnostic criteria. These have alternatively produced results in concordance with the initial findings, as in Rosenberg et al. which reported concordance of 88% in MZ versus 31% in DZ twins [23], while another study by Hallmayer et al. estimated a much lower heritability of 38%, in addition to identifying sex differences in concordance rates with female MZ concordance rate of 50% compared to 77% in males, and DZ concordance of 36% compared to 31% in males [24]. See [25] for a comprehensive review of these findings.

In summary, the quantitative findings, as well as indicating a strong genetic basis to the disorder, also support the notion of ASD as a spectrum [25]. They also hint at the underlying genetic architecture, being consistent with a multifactorial, threshold model [26], where genetic and environmental factors additively combine to increase liability to the point where it manifests as a clinically diagnosable disorder [27]. Within such a framework however, some of these findings become difficult to reconcile with each other. While the observed familial clustering supports the involvement of genetic or shared genetic and environmental factors [26], phenotypic variability and incomplete concordance in MZ twins are both highly suggestive of a role for non-shared environmental factors [28]. The implications are then that the underlying etiology of ASD cannot be completely explained by genetic factors, an idea important in the context of this thesis, and one which we shall explore further following a brief summary of the current molecular findings.

1.2.2 Molecular studies

Molecular genetics studies have revealed ASD to be highly genetically heterogeneous [29]. A number of syndromic or comorbid forms of autism are known to arise from large-scale chromosomal abnormalities or highly penetrant, single

gene mutations. As for non-syndromic, or idiopathic ASD cases, a number of risk conferring *de-novo* copy number variants (CNVs) and single nucleotide polymorphisms (SNPs) have been identified, many of which overlap with the risk loci identified for syndromic forms. We will now give a brief summary of some of the genetic findings and what they reveal about the etiology and pathomechanisms of the disorder.

1.2.2.1 Chromosomal abnormalities

Large, microscopically visible chromosomal abnormalities are observed in approximately 5-10 % of ASD cases [26], being predominantly associated with genetic syndromes showing comorbidity with autism. Some of the most recurrent findings are deletions and duplications in the long arm of chromosome 15, specifically the 15q11-q13 region [30]. A cluster of low-copy repeats in this region are particularly susceptible to re-arrangements, with paternal deletions being the primary cause of Prader-Willli syndrome, and maternal deletions giving rise to Angelman syndrome - both developmental conditions associated with autism [31]. Maternally inherited duplications are also frequently observed in the form of interstitial tandem repeats or as an extra chromosome with multiple copies, leading to the over-expression of genes which are ordinarily silenced (by imprinting - discussed later) [32], and causing moderate to severe autistic phenotypes [33]. Abnormalities in the 16p11.2 region represent another frequent finding, with deletion in this region resulting in severe, highly penetrant autistic phenotypes with dysmorphic features, and duplication of the same region leading to a less severe behavioural phenotype. Some further chromosomal abnormalities associated with autism include translocations at 7q22, and microdeletions in 22q11, specifically in the q11.2 and q11.3 regions - associated with rare conditions such as velocardiofacial syndrome, DiGeorge syndrome, and conotruncal anomaly face syndrome (see [33]).

1.2.2.2 Rare, single gene mutations

Monogenic forms of ASD where the primary genetic lesion is known account for 5% of ASD cases [34]. Fragile X is an X-chromosome linked form of mental retardation affecting males, where the underlying cause has been identified as a mutation to the *FMR1* gene. This leads to transcriptional silencing of the gene product *FMRP*, an RNA binding protein which is involved in regulating the translation of a number of other mRNAs at the synapse [35]. Rett syndrome is a progressive neurodevelopmental disorder affecting females where the primary genetic lesion is a loss of function mutation in the *MECP2* gene, coding for a methyl-CpG-binding protein important for regulating chromatin structure and levels of gene expression in neurons. Prader-Willli and Angelman syndromes arise from disruptions to imprinting in the 15q11-q13 region (previously implicated by cytogenic studies). Angelman syndrome has been linked to mutations in the *UBE3A* gene - which encodes a ubiquitin ligase that is important for protein degradation and is normally expressed exclusively from maternal alleles in neurons [36]. Macrocephaly is frequently comorbid with autism, and in a number of cases, mutations to *PTEN* have been identified - a phosphatase involved in cell cycle and apoptosis that is believed to regulate neuronal cell size through translational control [6].

A number of other rare mutations are also associated with ASD, which are typically observed in sporadic or familial cases. Identified genes include neuroligins *NLGN3/4* [37] and neurexins *NRXN1* [38] and *NRXN3*, which are involved in synaptic cell adhesion. Also recurrently implicated are genes coding for multiple ankyrin repeat domains such as *SHANK2* [39] and *SHANK3* [40] - scaffolding proteins that are important for synaptic organization, and other genes important for synaptic function including *CNTN3/4/5* [41] and *CNTNAP2* [42].

1.2.2.3 Copy number variants

CNVs describe submicroscopic chromosomal deletions and duplications ranging from around 1 Kb in size up to 1Mb, which are widespread throughout the human genome and are believed to contribute to genetic differences between individuals

to a similar extent as sequence variants [43]. Overall, *de-novo* CNVs are believed to be present in 10% of cases [44], but it is not currently known to what extent they contribute to overall risk. In terms of the loci identified, these often overlap with the chromosomal regions disrupted in the various comorbid autisms detailed previously, with risk-associated CNVs frequently identified in the 15q11-13 and 16p11.2 regions, and to a lesser extent in other regions such as 1q21, 5p15.2, 7q11.23, 17p11.2, and 22q11.2 (see [8] and [45] for recent reviews of findings). A number of CNVs have also been found either within or in the vicinity of some previously identified risk genes like *NLGN1* [46], *NRXN1* [44], SHANK1 [47], *CNTN4* [46], and *CDH8* [48].

1.2.2.4 Common variants

Genome wide association studies (GWAS) aim to uncover common, risk conferring variants for complex diseases. A small number of these studies have now been conducted for ASD. The Gene Discovery Project of Johns Hopkins and the Autism Consortium genotyped 3000 subjects from 780 multiplex families, and identified an associated SNP rs10513025 on chromosome 5p15 between the *SEMA5A* and *TAS2R1* genes [49]. The Autism Genome Project (AGP) Consortium performed a genome-wide scan of 4712 subjects from 1558 ASD families and identified rs4141463 in the *MACROD2* gene as the strongest signal [50]. Wang et al. genotyped 3103 subjects from 780 samples and identified a number of associations between the cadherin *CDH9/CDH10* genes, with the most significant hit being the SNP rs4307059 [51]. More recently, and using a much larger sample, the Cross-Disorder Group of the Psychiatric Genomics Consortium (PGC) analysed SNP data for 33,332 cases and 27,888 controls looking at five psychiatric disorders including ASD, and identified significant associations on chromosomal regions 3p21 and 10q24 and rs2799573 within the *CACNB2* gene (coding for calcium channel subunits).

1.2.2.5 Molecular mechanisms

Taken together the identified genes converge on a number of common pathways and cellular processes, implicating transcriptional dysregulation (as in both Fragile X and Rett), epigenetic dysregulation (e.g. disruption of parent-of-origin based expression patterns as in Prader-Willi and Angelman, *CDHs*), cell growth signalling (implicated in macrocephaly and autism), inhibition of PI3K/mTOR signalling (e.g. *PTEN*), synaptic adhesion/synaptic organisation (neurexins, neuroligins, *SHANKs*, *CNTNAPs*), synaptic function (*SCN1*, *SCN2*, *CACNB2*). Findings from functional studies further support a role for some of these processes in ASD pathogenesis, as we shall discuss further in the relevant sections to come.

1.2.2.6 Genetic architecture

The relative lack of success of GWAS in identifying common risk conferring variants for ASD has led some to question our assumptions about the likely genetic architecture. Added to this, the odds ratios for identified genes so far have been small and they have largely failed to replicate in other studies [52]. Broadly speaking, variance to ASD liability arises from common variation (according to the "common variant common disease" hypothesis [53]), rare variation ("rare variant common disease" [54]), or the environment -which can be shared and non-shared (described in further detail below). These are not mutually exclusive, and risk is likely to reside in each component - although the overall contribution of each of these remains a contentious issue.

Common variants: The common variant hypothesis posits that variants of small effect (1.1 – 1.5 fold) that are present at a high frequency in the population (> 5%), additively contribute to the variance of a trait [53, 55]. More generally in GWAS, there has been a relative lack of success in identifying these risk conferring variants, the so called "missing heritability" issue [55]. A number of methodological issues with GWAS can explain at least a portion of this missing heritability. Those commonly put forward include the use of small sample sizes underpowered to de-

tect common variants of small effect [55], causal variants not being in sufficient linkage disequilibrium (LD) with measured SNPs to be detectable by association [56, 57], and difficulties with disease diagnosis and ascertainment leading to sample heterogeneity or stratification (and a reduction in power) [58]. This last point could be particularly applicable to ASD, given the known phenotypic heterogeneity, with the various syndromes, comorbidities, and monogenic causes displaying different symptoms each displaying varying severities.

Several studies have attempted to measure the aggregate effect of common variants on ASD liability [59, 60, 61]. These use approaches such as Genetic Relatedness Estimation through Maximum Likelihood (GREML), which estimates total variance of liability by assessing the genetic relatedness of cases and controls across all the SNPs measured on a genotyping array, as implemented in the Genome-wide Complex Trait Analysis (GCTA) package [62]. In one of the first studies to apply the GREML approach to ASD, Klei et al. used 965 families from the Simons Simplex Collection (SSC) cohort and and 1141 families from the Autism Genome Project (AGP) cohort, to derive estimates for a narrow sense heritability of > 60% for multiplex families and 40% for simplex families [59]. A study by Cross et al. using the previously mentioned PGC sample investigated the heritability of five different psychiatric disorders (as well as the correlation between these), and obtained a figure of 17% for the heritability by SNPs. Finally, in the most recent study, Gaugler et al. [61] used the PAGES cohort (Population-Based Autism Genetics and Environment Study) with over 1.6 million families to derive a heritability estimate of ~ 52% [61].

Overall, while SNP-based estimates of ASD heritability vary, they are consistently lower than the previous estimates from twin studies. Since the GREML method is based on a different set of assumptions it produces a lower bound estimate of heritability, meaning that SNP-based estimates are not necessarily incompatible with twin-based estimates. The gap between these estimates could be the result of SNPs only tagging a proportion of the additive genetic variance. Alternatively, it has been suggested that initial twin-based estimates were upwardly biased due to various methodological issues including : inability to account for epistasis (G x G interactions) [63], violation of the equal environments assumption [64], and

failure to account for non-additive effects [61]. In support of the initial estimates being inflated, more recent twin studies have tended to produce estimates closer to that of the SNP-based estimates, with one of the largest and most recent studies by Hallmayer obtaining a value of 38% for heritability [24]. As well as re-igniting the debate about the extent to which inherited genetic risk plays a role in ASD, these SNP-based estimates of heritability are also taken as support of the theory that methodological issues are to blame for the lack of progress in identifying common variants in ASD GWAS, and offer some reassurance that increases in sample sizes should start yielding results.

Rare variants: Also contributing to disease risk are rare genetic variants present in low frequencies in the population, which account for a large proportion of individual risk [54, 43]. Support for a substantial rare inherited component to ASD comes from the observation that at least 10% of cases are the result of chromosomal abnormalities and highly penetrant mutations. Indeed, a number of of studies have attempted to quantify the contribution of rare inherited variants and CNVs to ASD susceptibility. Marshall et al. performed genome-wide assessment of CNVs and other structural abnormalities in 427 unrelated ASD cases, finding inherited CNVs in 196 cases [65]. In a study of 411 families with sporadic ASD, Krumm et al. found an increased burden of inherited CNVs in ASD cases compared to their unaffected siblings, which were transmitted preferentially from the mother [66].

In terms of the contribution from *de-novo* genetic point mutations and CNVs, there have been multiple studies that have attempted to quantify these. For example, study by Marshall et al. found that 7% of randomly selected idiopathic cases from simplex families harboured *de-novo* CNVs, of which 11% had two or more, while only 2% of cases from multiplex families had CNVs, compared to 1% in controls [65]. More recent studies have made similar findings, observing *de-novo* CNVs in approximately 8% of ASD cases compared to unaffected siblings [67, 68]. Another two studies examined the impact of heterozygous *de-novo* loss of function mutations, finding that these occurred in around 20% of cases compared to unaffected siblings [69, 70]. Gaugler et al. estimated the overall contribution of rare

de-novo events to ASD and arrived at a slightly lower figure of 3% [61]. Here it was also suggested that while *de-novo* mutation is likely to make a significant contribution to individual liability, contribution to overall heritability was not likely to be substantial.

Environment: Within the twin modelling paradigm, environmental influences can be divided into shared/ common (C) - those affecting both twins equally and serving to make them more phenotypically similar, and non-shared/ unique (E) - which make twins growing up in the same family different. The non-shared environment comprises *de-novo* genetic as well as non-genetic effects, the later of which can include epigenetic, gene expression, environmental, or stochastic factors [25]. In ASD, it has been suggested that gene regulatory effects - potentially mediated by epigenetic mechanisms, could be driving MZ discordance [24, 28]. While such factors do not contribute to the heritable component of the disorder, they could account for a substantial component of risk and perhaps paradoxically, they could also explain some missing heritability. One way in which they could contribute to missing heritability is by inflating concordance rates through "false heritability" [71] and incorrectly partitioning non-additive genetic, *de-novo* variation, and gene-environment interaction (G × E) into the additive genetic component [72]. Related to this, there is the possibility that gene-environment interactions could modulate risk by altering response to certain environments (phenotypic plasticity), which could further mask true genetic associations. More recent heritability estimates leave plenty of room for the environment, for example, Gaugler et al. estimated the contribution of the environment to ASD liability to be approximately 40%, split between common and unique sources [61].

1.2.3 Conclusions

Overall, the findings from quantitative and molecular studies support a genetic basis for ASD [25]. However, currently the identified common and rare variants and *de-novo* mutations explain no more than 20% of cases [11], and related to this, there is still no coherent explanation of the underlying pathomechanisms of the

disorder. The lack of progress thus far in uncovering the genetic risk architecture of ASD is slightly perplexing, given that early estimates of the narrow sense heritability for ASD placed this as high 90 % [22]. More recent studies which use SNP-based methods to estimate heritability suggest values of closer to 52% [61]. It would seem reasonable to deduce then that in the case of ASD, additive genetic risk is likely to be a much smaller piece of the complete etiological puzzle than initially anticipated, and to better understand the biological basis of the disorder, we must also investigate the impact of non-additive genetic and environmental factors.

In terms of where exactly this "dark matter" of liability lies, comparisons of MZ and DZ twins have consistently found a moderate contribution from non-shared environment, with estimates placing this between 20% and 38% [73, 74, 75, 76]. Epidemiological investigation of non-shared environmental factors is however fraught with difficulty, especially given that these are likely to occur early in development, and also because of the potential for confounding and reverse causation. This is where investigation of intermediate molecular layers such as gene expression and epigenetics may prove fruitful, as these are more tangibly linked to genetics as well as underlying cellular processes, and can also reflect environmental influences, potentially persisting long after the initial insult. In this thesis, we submit that the characterisation of molecular differences between ASD discordant MZ twins could provide a powerful means to uncover the biological mechanisms underlying the non shared environment in ASD. Before we review the supporting evidence for this, we will first cover some necessary background on epigenetics and regulation of gene expression.

1.3 Epigenetic regulation of gene expression

1.3.1 Epigenetics

The term epigenetics is used to refer to mitotically heritable chemical modifications to DNA that alter the expression of a gene without changing the underlying

sequence [77, 78]. Epigenetic regulation of gene expression is crucial for establishing cell-specific patterns of expression during normal development [77, 78], and for gene silencing - to control gene dosage in females by X-inactivation [79, 80], or for allelic-silencing at specific loci according to parent-of-origin during imprinting [81, 80].

A variety of different epigenetic modifications and marks act in concert to either promote or repress a transcriptionally active chromatin state. These primarily consist of chemical modifications to DNA such as cytosine methylation (discussed below) and hydroxymethylation - which block transcription factor binding or recruit chromatin structure modifying proteins, and post-translational histone tail modifications - which directly alter chromatin conformation. Examples of histone tail modifications include: the addition/removal of methylation groups from lysine residues (e.g. methylation of H3K4 and H3K36 for transcriptional activation, and of H3K9 and H3K27 for repression), and also of acetyl groups (e.g. acetylation of H3K9 and H4K14 for activation) - see [82] for an in-depth review. In addition to these, there are a number of non-coding RNAs (ncRNAs) that interact with the above mechanisms and also play an integral part of the epigenome (discussed further in the next section).

1.3.2 DNA methylation

DNA methylation is one of the best characterised epigenetic marks, describing the addition of a methyl group to the 5 carbon of a cytosine base to form 5 methylcytosine (5mC); a reaction catalysed by a group of enzymes called DNA methyltransferases [78]. This predominantly occurs in the context of CpG dinucleotides, the majority of which are methylated. There are around 28 million such sites in the human genome [83], and these tend to cluster in regions referred to as CpG islands, characterised by high GC and CpG content [84]. These regions are typically associated with gene promoters, and are usually unmethylated, indicating a transcriptionally active state for the gene, becoming methylated in the process of differentiation [77]. DNA methylation of CpG islands promoters results in downregulation of gene expression, typically by blocking transcription factor binding,

or by attracting methyl-binding proteins that in turn recruit chromatin modifying activities [77]. CpG sites are also present downstream of promoters in gene bodies, in both island and non-island contexts, where the relationship between methylation and transcriptional regulation is not as generalizable and varies according to genomic context, either following the canonical mechanism for silencing and/or downregulation, or paradoxically resulting in upregulation of expression [85].

1.3.2.1 Developmental establishment of methylation marks

During development, DNA methylation and histone marks are erased during two distinct phases of epigenetic reprogramming (see review articles: [86, 87]). This resetting of the epigenome occurs in primordial germ cells from day 10.5 to 13.4, and then in the zygote after fertilization and extending into preimplantation development [86]. The erasure of DNA methylation is widespread and comprehensive, with overall methylation levels dropping as low as 10% (compared to 70% prior) [88], thus allowing *de-novo* establishment of methylation patterns during development to occur on a "blank slate". These new methylation patterns are established first in the germ cells, targeting CGIs which are required to be methylated, for example, imprinting control regions which are important establishing parent-of-origin specific expression in imprinted genes. A second wave of erasure then occurs in the zygote post-fertilization in distinctive patterns on the maternal and paternal genomes, this time not including imprinted regions. Following blastocyst implantation, new methylation marks are then laid down which are important for cell differentiation [87].

1.3.3 Methylomic assays and technology

1.3.3.1 Bisulfite sequencing

A number of different experimental assays have been developed to measure DNA methylation status. The assay used largely depends on the study design and the

aims of the study. For in-depth profiling of methylation status for a sequence of interest in a small number of samples, the gold standard is whole-genome bisulfite sequencing (BS-seq). BS-seq involves direct sequencing of DNA which has undergone treatment with bisulfite, which converts unmethylated cytosine bases to uracil (subsequently amplified as thymine during PCR amplification), whilst leaving methylated cytosines unconverted [89]. Average, per-site methylation levels can then be determined by effectively performing SNP calling at the sites of interest, and quantifying the proportion of those which have been unconverted to those which were converted. BS-seq can in theory be used to profile all 28 million sites, but in reality such an approach would be prohibitive in terms of both time and cost for studies involving a large number of samples [83].

1.3.3.2 Microarray

Epigenome-wide association studies (EWAS) attempt to characterise patterns of methylation genome-wide, allowing for a hypothesis-agnostic investigation into methylation for the trait or condition under consideration. For these types of study, array based platforms are the most popular as they offer a high-throughput quantification method, which sacrifices coverage and measurement accuracy for reduced cost and ease of use [90]. The Illumina 450K array is one of the most widely used, profiling just over 485,500 individual CpG sites across the genome, covering <2% of all known CpG sites, and 99% of RefSeq genes [91]. Briefly, the array technology works as for gene expression measurement (see next section), but in this case measuring methylation of CpG sites by detecting base changes induced in a sample following bisulfite conversion, by using probes with complimentary sequence for the methylated and unmethylated variants of the alleles, yielding a quantitative measurement of the proportion of methylated CpGs present at each target sequence [92].

1.3.4 Epigenetic mechanisms in disease

There has been particular interest in the role of epigenetic mechanisms in complex disease, perhaps motivated by a number of important observations. Firstly, there is the fact that marks are established early during development and undergo two major waves of reprogramming. These periods of epigenetic plasticity represent critical periods where dysregulation could, for example in the case of neurodevelopmental disorder, disrupt brain development with profound long-lasting implications for disease susceptibility in later life (this might be considered a neurological re-contextualising of the "developmental origins" hypothesis [93]). Secondly, numerous studies have shown that CpG methylation in particular, as well as displaying inter-tissue variation within individuals, is also variable between individuals [94]. It has therefore been suggested that methylomic variation could account for some portion of phenotypic variance currently attributed to non-genetic factors, and represent a plausible mechanism through which environment can interact with genes. The "common disease genetic and epigenetic" (CDGE) hypothesis [95], extends the common disease model to include this epigenetic component. Such a model could help to explain other complex disease phenomena, for example, gender-based differences [96], MZ twin discordance [97], and age-related increase in incidence [97].

The role of methylomic variation has been investigated in many conditions such as cancers [98, 99], autoimmune disorders [100, 101] and neurodevelopmental conditions[102, 103, 104]. A number of studies have looked for methylomic signatures in ASD, and these will be summarised briefly in section 1.5.1.

1.4 Profiling patterns of gene expression

1.4.1 The transcriptome

The transcriptome is one of the most fundamental molecular levels at which phenotype can be influenced. As well as varying across tissues types, dynamically

throughout the cell cycle (and over the course of development), gene expression also displays natural variation between individuals [105]. Since gene expression is only moderately heritable (~30%) [106, 107, 108], these differences could then reflect genetic as well as environmental, epigenetic, or stochastic differences. This last point helps explain the current revival in transcriptomic studies, given the role of the transcriptome as an intermediate, molecular phenotype, which also potentially links genes and environmental effects via epigenetic mechanisms.

1.4.2 RNA varieties

Gene expression typically refers to protein coding messenger RNAs (mRNAs), which have important functional roles in the cell, and influence cellular phenotype in tangible ways. One of the more surprising findings from contemporary genomics, however, has been that despite the fact that the majority of the genome appears to be transcriptionally active, only around 2% of genes are actually protein coding [109]. In addition to important infrastructural RNAs (e.g. tRNA), are a variety non-coding RNAs (ncRNAs), that are believed to be important for gene regulation [110]. Small ncRNAs include small interfering RNAs (siRNAs), micro RNAs (miRNAs), and PIWI-interacting RNAs (piRNAs), which regulate the activity of of mRNA primarily through RNA interference. Long non-coding RNAs (lncRNAs), are also involved in regulation, in some cases via interaction with epigenetic mechanisms such as chromatin remodelling and histone tail modification [111]. A well characterized example of this type of regulation in action is that of the lncRNA *XIST*, which silences the inactive X chromosome during X-inactivation by preventing the transcriptional machinery from accessing the region and additionally recruiting chromatin remodelling complexes to promote an inhibitory conformation [112].

1.4.3 Expression profiling assays and technology

1.4.3.1 Microarray

For well over a decade, oligonucleotide arrays have been the technology of choice for expression profiling [113]. Microarrays consist of millions of short fragments of known DNA sequence corresponding to partial sequences from annotated genes (called probes), which are arranged and fixed onto a solid surface, typically a glass slide. Transcripts are isolated from the sample, treated and then reverse transcribed into cDNA, which is fluorescently labelled and hybridised to the array - these constitute the target sequences. Any targets with sequence complementary to probes will bind causing fluorescence to be emitted with intensity proportional to the the abundance of mRNA. The arrays are imaged in order to detect these binding events, yielding intensity readings that can then be quantified and used to derive gene expression estimates (see [114] and [115] for detailed descriptions of the technology).

1.4.3.2 RNA sequencing

In the past few years, microarrays have began to be replaced by RNA-sequencing (RNA-seq) technology for gene expression profiling. In an RNA-seq experiment, next generation sequencing is used to determine the sequence of bases making up the various RNAs contained within in the sample (both mRNA and non-coding varieties). In general, an RNA-seq experiment proceeds as follows: RNA from the sample (either mRNA or total) is isolated, fragmented, and then converted to cDNA libraries with adaptor sequences bound at both ends, which are then amplified by PCR and then sequenced using high-throughput genomics sequencing technology. The resulting sequence reads are then mapped back to the genome and quantified to produce a quantitative measure of gene expression (see [116] for a summary). Utilising RNA-seq allows for a higher resolution investigation to be conducted, enabling detection of a much wider range of gene expression including low abundance transcripts that might otherwise be undetectable by microarray

[117, 118], and non-coding varieties such as lncRNAs [119]. Because information about the actual sequence is provided, other regulatory phenomena can also be investigated, including alternative splicing [118], and parent-of-origin effects (which results in allele specific expression). Stranded approaches which preserve information on which strand was transcribed form additionally allow antisense transcripts to be detected [120]. Antisense transcripts are also believed to play an important role gene regulation - a well-known example being the inhibitory role of the *TSIX*, the antisense complement of the lncRNA *XIST*, which acts to regulate the expression of *XIST* during X-chromosome inactivation [112].

The main disadvantages of RNA-seq are the much greater cost associated with sequencing experiments, and a higher barrier for entry because of the computing resources and bioinformatics required to analyse the data. Proclamations about the "death of the microarray" may have therefore been premature, as for certain applications, for instance smaller scale exploratory gene expression studies, microarrays can still often be the preferred option.

1.5 Non-genetic factors in ASD

1.5.1 Epigenetic

Current interest in the potential involvement of epigenetic mechanisms in ASD is largely motivated by 1) demonstration of gene dosage effects for example at the 16p11.2 region [121], 2) the disruption of imprinting mechanisms observed in syndromic forms of ASD such as Prader-Willi and Angelman, and 3) the involvement of methylomic mechanisms in the pathologies of Fragile X and Rett syndromes. In the case of Fragile X, the FMR1 gene contains a polymorphic sequence CCG in the 5'-UTR which can undergo triplet repeat expansion, leading to hypermethylation of the promoter, histone deacetylation, and silencing of the gene [122]. FMR1 is itself encodes an RNA-binding protein, helping to regulate the expression of a number of other mRNAs at the synapse, and its loss therefore impacts the normal functioning of the synapse [35]. The *FMR1* gene has also been implicated in idio-

pathic ASD, with decreased gene expression in ASD affected individuals observed [123]. As for Rett syndrome, the majority of cases are caused by a mutation in the *MECP2* gene, which encodes a methyl-CpG-binding protein important for transcriptional regulation, histone modification, and alternative splicing at the synapse. Mutations in the *MECP2* gene have been observed in a small number of ASD cases [124], as has familial transmission of *MECP2* variants [125], and altered gene expression in ASD affected individuals as compared to controls [123]. As discussed earlier, Prader-Willi and Angleman syndromes both arise from genetic lesions in the 15q11-q13 region, with large paternal deletions giving rise to Prader-Willi and large maternal deletions *UBE3A* giving rise to Angelman. This region is also subject to genomic imprinting, and it has been demonstrated that deletion of the identified imprinting control regions can give rise to both Prader-Willi and Angelman [126]. In the case of Angelman, the causative gene has been identified as *UBE3A*, where loss of function mutations cause the phenotype when inherited maternally, but are benign when inherited paternally [126]. This is because expression of the gene is tightly regulated via epigenetic mechanisms in a parent-of-origin specific manner, with the paternal allele ordinarily silenced and only the maternal allele expressed.

Motivated by these observations, there have been at least two recent investigations into patterns of genome-wide methylation in ASD. A study by Nguyen et al. profiled methylation in lymphoblastoid cell lines derived from 3 MZ discordant twin pairs, 2 unaffected siblings, and 1 concordant control pair [127]. Autistic pairs were diagnosed using the autism diagnostic interview-revised (ADIR) instrument, and methylation profiling performed on CpG island arrays. The results revealed 201 genes with evidence of differential methylation when profiles from both co-twins were compared to their unaffected siblings. Potentially relevant findings included the identification of *BCL-2* and *RORA* - which both displayed promoter hypermethylation, and pathway analysis revealed enrichment for pathways involved in transcriptional control of expression, and nervous system development and function. Another study by our group examined methylation of 50 MZ twins pairs including concordant, discordant, trait discordant, and age-matched controls, and identified a number of differentially methylated regions associated with ASD [128].

As this is highly relevant to the work described here, we reserve further discussion of the findings until Chapter 2.

1.5.2 Environmental

It has been suggested that environmental factors could trigger the development of ASD in individuals with an underlying genetic predisposition [129, 28]. Indeed, a number of environmental risk factors for autism have been reliably replicated, including: prenatal viral infection (by influenza, rubella, and cytomegalovirus [130]), zinc deficiency [131], prenatal exposure to toxins (valproic acid [132], thalidomide [133], psychiatric drugs [134]), and paternal age [135] (see [136] for an in-depth review of the published findings).

Given that the timing of many of these exposures is early in fetal life, one suggested mechanism by which ASD risk might be increased is through epigenomic disruption and altered gene expression patterns during critical periods of brain development [136, 28]. Support for this comes from the recognition that a number of prenatal environmental risk factors are likely to directly interact with epigenetic marks. For example, valproic acid is an anticonvulsant drug and is commonly prescribed for treatment of seizures and mental health disorders such as bipolar disorder. Administration of valproic acid during pregnancy has been shown to increase the risk of ASD by up to 16 times that seen in the general population [132], and studies in rodents have shown also shown that administration *in utero* can impact neurodevelopment leading to deficits in social behaviour and anxiety [137]. Given that the agent is a known histone deacetylase inhibitor, one potential mechanism of action is the disruption to methylation patterns and global levels of gene expression [138].

A number of studies have implicated levels of folate (vitamin B9) micronutrient supplementation during pregnancy as a risk factor [139, 140, 141, 142].This remains something of a controversial topic, especially since some of the findings seem to be contradictory [139], with studies alternatively finding that an insufficient [141, 142] or excess [140] folate increases ASD risk. While clearly further

research is required to establish the veracity of this link, there is at least some level of biological plausibility. To begin with, periconceptional folate supplementation has been shown to protect against neural tube defects [143]. One proposed mechanism of action for this protective effect is through involvement of folate as a methyl donor in the one carbon metabolic pathway [144, 145]. This pathway is essential for the establishment of DNA methylation patterns, and disruption has been shown to decrease global methylation and affect gene expression, leading to abnormal brain development [145]. Linking this to ASD, are associations of mutations in genes coding for constituents of the one carbon pathway, specifically S-adenosylmethionine *SAM* [146] and 5-methyltetrahydrofolatereductase *MTHFR* [147, 148].

Revising the multifactorial model: Based on the various genetic and epidemiological findings, it seems likely that ASD susceptibility arises from both common and rare variation, as well as epigenetic and environmental factors. An important challenge is combining these in a unified model for population risk. While no such model currently exists, there have been a number of attempts to combine genetic and epigenetic effects and inherited and *de-novo* mutation for particular subsets of ASD cases. Jiang et al. proposed a oligogenic interaction model, where a combination of genetic and epigenetic factors both *de-novo* and inherited affect the expression of a small number of causative genes. This model is based on the epigenetic and parent-of-origin effects observed in Prader-Willi and Angelman syndromes, but the authors speculate that it might be more generally applicable in ASD [126]. Zhao et al. argued for a two-class risk model, where the majority of ASD arises from *de-novo* mutation, arising in the parental germ line. Female offspring have some protection against the behavioural manifestations of such mutations, but can then carry the mutation, which is then passed on in a dominant mode of inheritance [149].

While risk modelling is outside of the scope of this thesis, should non-genetic effects be found to have a substantial influence on ASD trait liability, then the multifactorial threshold model will need to be revised. A suitable model would have to account for both genetic and epigentic risk, as well as being able to account for the

heritability, phenotypic variance, and gender bias observed in ASD, as well as the frequent co-occurrence with rare monogenic syndromes [150].

1.6 Gene expression profiling in ASD

1.6.1 Systematic review of published findings

It is perhaps not surprising given the importance of the transcriptome as an integrator of genetic, epigenetic and environmental influences, and the evidence for epigenetic dysregulation in monogenic and comorbid forms of autism, that a number of previous studies have investigated gene expression in ASD. To identify relevant studies, a Pubmed search was conducted using the terms: (gene expression[Title] OR transcription[Title] OR transcriptomic[Title]) AND (autism[Title] OR autistic[Title] OR ASD[Title]) AND "humans"[MeSH Terms]. This yielded 47 results, which were then manually filtered to retain only those using genome-wide approaches, giving a final list of 15 papers (shown in Table 1.1). Here we summarise the findings.

Reference	Sample	Gender (M:F)	Age range	Tissue	Platform	Main findings
Purcell et al. (2001)	10 ASD 23 controls	9:1	5-54	Brain (CB,PFC,CN)	Clontech Atlas Human Neurobiology (MA)	↑ glutamate/AMPA
Hu et al.(2006)	6 ASD/non (MZd) 4 ASD (MZc)	6:0 4:0	6-16 6-16	LCL	TIGR 40K (MA)	↑ neurological function ↓ 5-HTT ↑↓ nervous system development ↑ inflammation
Nishimura et al.(2007)	8 ASD/FX 7 ASD/15q11-q13 15 controls	8:0 7:0	N/A	LCL	Agilent WHG G4112A (MA)	↑↓ cell communication ↑↓ immune related ↑ JAKMIP1 ↑↓GPR155
Garbett et al. (2008)	6 ASD 6 controls	4:2	4-30	STG	Affymetrix U133 Plus 2.0 (MA)	↑↓ cell communication ↑↓ differentiation ↑↓ cell cycle ↑ immune related ↓ neuronal development ↑ transcriptome variability
Gregg et al.(2008)	35 AU 14 ASD 12 controls	30:5 13:1	N/A	Blood (whole)	Affymetrix U133 Plus 2.0 (MA)	↑ NK/CD8+ cell mediated cytotoxicity ↑↓ glutamate neurotransmission
Enstrom et al.(2009)	35 ASD 12 controls	30:5	2-6	Blood (whole)	Affymetrix U133 Plus 2.0 (MA)	↑ NK ↓ cellular proliferation
Hu et al.(2009)	21 ASD 21 unaff sibs	21:0	4-14	LCL	Invitrogen custom (MA)	↑↓ cholesterol metabolism ↑↓ androgen biosynthesis ↑↓ immune related ↑↓ nervous system development ↑↓ cytoskeleton
Hu et al.(2009b)	86 ASD 30 controls	86:0	5-28	LCL	TIGR 40K (MA)	↑↓ circadian rhythm ↑↓ androgen sensitivity ↑↓ inflammatory response ↑↓ axon guidance ↑↓ apoptosis ↑↓ neurogenesis ↑↓ cell survival ↑↓ transcriptional regulation
Alter et al. (2011)	82 ASD 64 controls	82:0	mean = 5	Blood (PBL)	Affymetrix HuGe U133 Plus 2.0 (MA)	↓ variance global expression ↓ transcriptional regulation ↑↓ zinc
Seno et al (2011)	20 ASD 22 unaff sibs	13:7	N/A	LCL	Illumina Human-Ref8 v3 (MA)	↑↓ nervous system development ↑↓ long term potentiation ↑↓ NOTCH signalling ↑↓ neuroligical disorders (RETT)
Kuwano et al. (2011)	21 ASD 21 controls	17:4	18-35	Blood (PL)	Agilent WHG oligoDNA (MA)	↓ ribosomal ↑ MECP2 ↓ FMR1
Voineagu et al. (2011)	19 ASD 17 controls	14:5	2-56	Brain (STG,BA9,CV)	Illumina Ref v3 (MA)	↓ synaptic ↑ immune/inflammatory ◊ A2BP1/FOX1
Chow et al. (2012)	9 ASD (child) 7 control (child) 6 ASD (adult) 11 control (adult)	9:0 6:0	2-14 15-56	Brain (DLPFC)	Illumina Human-Ref8 v3 (MA)	↑↓ neurogenesis ↑↓ cell cycle DNA-damage resp ↑↓ apoptosis and survival ↑↓ immune response ↓ cytoskeleton and ECM ↑↓ A2A receptor ↑↓ development ↑↓ signalling ↑↓ oxidative stress ↑↓ WNT
Glatt et al. (2012)	37 ASD 23 PDD 17 DD 34 LD 27 controls	29:8	1-3	Blood (PBMC)	Illumina WG-6 v3 (MA)	↑↓ immune response ↑↓ globin
Kong et al. (2013)	20 ASD 20 unaff sibs 18 controls	17:3	4-17	Blood (whole)	Affymetrix Gene 1.0 ST (MA)	↑ ribosomal ↑ spliceosomal ↑ mitochondrial ↓ neuroreceptor-ligand ↓ immune response ↓ calcium signalling

AD = autistic spectrum disorder
PDD = pervasive developmental disorder
DD = developmental delay
LD = language delay
MZd= MZ discordant
MZc = MZ concordant
FX = Fragile X

PL = peripheral leukocytes
PBMC = peripheral blood mononuclear cells
STG = superior temporal gyrus
BA9 = prefrontal cortex
CV = cerebellar vermis
DLPFC = dorsolateral prefrontal cortex
PBL = peripheral blood lymphocytes
LCL = lymphoblastoid cell line
NK = natural killer cells
CB = cerebellum
PFC = prefrontal cortex
CN = caudate nucleus

MA = microarray

↑ = up regulated
↓ = down regulated
↑↓ = differentially regulated
◊ = differentially spliced

Table 1.1

In one of the first studies to investigate gene expression in the context of ASD, Purcell et al. examined global gene expression changes in post mortem cerebellar tissue for a sample of 10 ASD affected and 23 control subjects [151]. Two different early array platforms were used (covering < 10,000 genes), with polymerase chain reaction (PCR) used for validation of top differentially expressed genes, and Western blotting used as a complementary approach to measure the protein levels for these genes. Following standard data processing steps, differential expression between the groups was assessed using t tests. Overall, while the majority of genes were found to show no differences in expression between ASD and control, 30 genes were found to be differentially expressed, including genes related to glutamate neurotransmission: *GluR1,GluR2, GluR3* and *EAAT1, EAAT2*, which were found to be upregulated in ASD cases compared to controls. Protein levels for *GluR1,EAAT1, EAAT2* were also consequently found to be increased. In the discussion, the authors suggest that the results support a glutamate hypothesis of ASD, where given its importance for synaptogenesis and neuronal outgrowth during development, that it could be particularly vulnerable to disruption. Several limitations of the study are also acknowledged, including: the use of post–mortem tissues collected long after the onset of the disorder - making it entirely possible that the expression differences are secondary to the disorder, and the presence of sample heterogeneity – with clinical variations such as intelligence quotient (IQ) and presence of seizures.

Hu et al. performed genome-wide gene expression profiling of lymphoblastoid cell lines derived from 3 pairs of MZ twin pairs discordant for ASD and 4 pairs concordant for ASD [152]. For three sets of male twins, one member of each pair had met the ADI diagnostic criteria, while the other was not and classed as either broad spectrum or not quite autistic. Two pairs of concordant ASD twins were also used. Expression profiling was carried out on an array platform measuring ~ 40,000 genes, and the data processed. Differential expression analysis was performed using a modified t statistic test (SAM – significance analysis of microarrays), with significance thresholds set at FDR < 0.26 and log fold change of ~ 0.58). The results from the within-discordant pairs analysis showed 25 genes as being significantly up-regulated in the affected twin compared to the unaffected

twin, and 19 genes down-regulated (for the same thresholds). A number of these genes were found to correspond to genes involved in neurological development, and of those confirmed by qPCR included: *ASS, DAPK1, IL6ST, EIFC2*. Pathway analysis across all pairs additionally highlighted nervous system development and cytokine signalling as being enriched. Concluding, the authors suggest that their study supports the use of peripheral tissue for identifying biomarkers, and suggest that a blood based screen could be developed. In terms of the identified genes, it is also observed that a number of the neurologically relevant genes are also linked to inflammatory networks, consistent with the hypothesis of neural inflammation being important in ASD.

Another study by Nishimura et al. also profiled lymphoblastoid cell lines [153]. Here, the cell lines were derived from individuals with two monogenic ASD etiologies, with 8 males diagnosed as having Fragile X (FX) with associated autism and 7 males with a duplication in the 15q11-q13 region (dup (15q)), and 15 non-autistic males used as controls. Expression profiles were generated using microarrays, the data processed, and then compared across all three groups. A total of 293 genes were found to be differentially expressed using ANOVA and an FDR threshold of 5%, these included *FMR1* and *UBE3A* as expected (the single gene causes identified for these syndromes) and *CYFIP1* – a known antagonist of the FMR1 gene product. Clustering was performed using the DE genes, which successfully classified the individuals into their correct experimental groups. Further analyses identified an overlapping set of 68 genes that were dysregulated in both FX and dup (15q) cases, with 52 dysregulated only in FC and 12 in dup(15q). Validation of the expression of 19 selected genes was performed using qRTPCR, of which 17 were found to be differentially expressed. Pathway analysis revealed that genes involved in cell communication were upregulated in the 68 overlapping genes, along with enrichment of immune response. In the FX dysregulated gene lists, there was enrichment of chaperone and protein folding genes, and in the dup (15q) genes, there was enrichment for those involved in RNA binding and mRNA metabolism. Next, using a neuronal cell line and short hairpin RNA (shRNA), the expression of *FMR1* was reduced and *CYFIP1* to examine the the downstream effects. Two downstream genes were found to be significantly differentially expressed: *JAKMIP1* and *GPR155*, and

this was further validated in knock out mouse models. Finally, the expression of these two genes was examined 27 male sibling pairs discordant for ASD, where they were also found to be differentially expressed. Overall, the authors claimed the results firstly demonstrated that it was possible to distinguish ASD etiologies from each other based on gene expression signatures, secondly that there were likely to be commonly dysregulated pathways, and thirdly that the findings were generalizable to idiopathic ASD.

Expression profiling of post mortem temporal cortex in ASD was performed by Garbett et al. [154]. The samples were taken from the superior temporal gyrus of 6 ASD cases and 6 matched controls, gene expression quantified by microarray (interrogating approximately 38,000 genes), and the data processed following the standard analysis pipeline. Using paired t tests and significance cutoffs of log FC > 2 and $p < 0.05$, 152 genes were identified as differentially expressed of which 130 showed increased expression and 22 decreased expression. For validation, 20 genes were selected for qPCR analysis, in which 80% remained significant, including the genes *HSPA6, HSPB8, MAP2K3, NOTCH2* amongst others. Higher level analysis using gene set enrichment (GSEA) testing revealed significant overrepresentation for a number of immune system related pathways including those involved in antigen-specific immune response (e.g. *TOLL, TNFR2, IL2R*), inflammation (e.g. *NFKB, IL1R, GSK3*) , cell death (e.g. *NFKB, TNFR2, P38MAPK*), and auto-immune disease (e.g. *NFKB, TOB1, FAS*). In concluding, the authors suggest that the results firstly indicate that transcript activation is more common than repression in ASD cases, the autistic transcriptome is characterised by an over active immune response, that transcription of genes related to cell communication, differentiation and cell cycle regulation is altered in a potentially immune system-dependent manner, and that the transcriptome showed increased variability in ASD.

In one of the first studies to investigate the autistic transcriptome in whole blood, Gregg et al. [155] conducted microarray expression profiling using samples taken from 35 children with autism (clinically diagnosed using ADI-R and ADOS screening tests), 14 with ASD (not meeting full criteria in one or more domains), and 12 age matched and gender matched controls from the general population. Data was

processed as per standard procedures, and genes were tested for differential expression between the groups using unpaired t-tests, with FDR < 5% and log FC ≥ 1.5 used as thresholds. For the ASD group, no genes were found to be significantly DE when compared to either the autism group or controls. For the autism group, 55 genes were identified as DE when compared to the control group. The differential expression for a number of these genes was subsequently confirmed using RT-PCR. Pathways analysis revealed an enrichment of genes involved in natural killer (NK) cell-mediated cytotoxicity, including the *IL2RB* and *EAT2*. Overall, these results suggested that gene expression differences associated with autism are detectable in whole blood, and further taken as supporting a role for abnormalities in peripheral blood leukocytes in ASD.

Enstrom et al. also performed microarray expression profiling using whole blood samples obtained from 52 children with ASD and 27 typically developing controls [156]. This study focused primarily on the NK cell related genes that were previously identified in the Gregg et al. study [155]. The data was processed, and differential expression analysis was then performed using unpaired t-tests which identified 626 probes, 544 of which were downregulated and 82 upregulated in ASD compared to controls. From the list of upregulated probes, corresponding to 59 genes, 30 of these were associated with leukocytes and 22 were associated with cytolytic cells, mainly NK cells. To further investigate the functional and physiological implications of this altered gene expression, a number of single cell and direct protein analyses were performed. Cell isolation showed a significantly higher proportion of NK cells in the peripheral samples from the ASD cases, and an NK cytotoxicity assay, showed decreased NK activity in ASD. Concluding, the authors suggest that because NK cells are important for host defence against infections, play a role in autoimmunity and inflammatory processes in the brain, and are the predominant immune cells in early development, that altered NK cell response could modify susceptibility to infections potentially impacting neural development. Thus, differences in gene expression of NK cell related genes and the functional implications of these differences represent a neuroimmune explanation for a component of ASD risk.

Hu et al. interrogated genome-wide gene expression in lymphoblastoid cell lines

derived from 21 sibling pairs in which one sibling had diagnosed ASD and the other was non-autistic [157]. To reduce phenotypic heterogeneity, only ASD presenting with severe language impairment (as assessed by the ADIR screening test) were selected for the study. Expression profiles were generated using a microarray measuring ~ 40,000 genes, and following standard data processing steps, differential expression analysis was performed to compare ASD affected siblings with their non-autistic siblings, using a modified t statistic test (SAM) with significance threshold set at a log fold change of ~ 0.3. A number of differentially expressed genes of potential relevance were identified, including *SCARB1* and *SRD5A1*, both involved in steroid hormone bionsynthesis, and *SCN5A*, a sodium channel expressed in the limbic brain. Pathway analysis of the gene list revealed an enrichment of pathways involved in endocrine system development, cholesterol metabolism apoptosis, differentiation, inflammation, and epilepsy. A selection of 6 genes had their differential expression confirmed by qRT-PCR, and steroid profiling was performed on 3 randomly selected sib pairs, to test the hypothesis that the qRT-PCR confirmed differences in *SCARB1* and *SRD5A1* expression could result in increased testosterone biosynthesis, which was subsequently found to be the case. Overall, the results were taken to suggest that disruption to normal androgen levels could have consequences for neurodevelopment and function, potentially giving rise to the autistic phenotype (and perhaps explaining the gender bias).

Another study by Hu et al. [157] used lymphoblastoid cell lines derived from 116 individuals with ASD and age-matched, unaffected controls. ASD cases were stratified into 3 different groups according to severity of symptoms across the items in the ADIR screening test, and females excluded, along with individuals with known genetic or chromosomal abnormalities, those born prematurely, and those with comorbid psychiatric conditions. In an accompanying study, the resulting subgroups were analysed and found to correspond to mild ASD, severe language impaired, and ASD with savant skills endophenotypes. Gene expression profiles were generated using microarray technology to assess ~ 40,000 mRNAs across the genome, and standard array data processing steps were carried out. Differential expression analysis was then performed using SAM, with a two-class comparison of ASD affected and controls, and a four-class comparison of the three different ASD

subgroups and controls, with the significance threshold specified at FDR < 5%. Selected significant results were subsequently confirmed using qRT-PCR. The results revealed a set of significantly differentially expressed genes common across all ASD cases when compared to controls, including *ITGAM, NFKB1, RHOA, SLIT2, MBD2*, with some genes further showing a quantitative relationship between level of gene expression and severity of ASD phenotype. The analysis of the different subgroups revealed differentially expressed genes specific to those groups. Pathway analysis of the common DE genes in ASD revealed an enrichment of genes involved in synaptic transmission, neurogenesis, neurulation, long term potentiation, and protein ubiquitination. Separate pathway analyses were also carried out for the ASD subgroup. For the severely-language impaired group there was an enrichment for genes involved in apoptosis, in the mild ASD group there was enrichment for small molecule biochemistry, free radical scavenging, and cellular function, and finally in the savant group, RNA posttranscriptional modification was highly significant. In the severely-language impaired group, 15 genes involved in regulation of circadian rhythm were also identified. Next, to test whether the identified differentially expressed genes had any relationship with known ASD susceptibility loci, they were mapped to previously reported QTLs. It was found that around a third of significant genes were associated with autism QTLs, and that these were significantly enriched on chromosomes 7, 10, 16, 17 and 22 for the combined group. Concluding, the results showed differences in gene expression between the ASD subgroups and controls, with specific pathways implicated in the different subtypes, and common genes implicated across subgroups indicating that basic deficits may also underlie ASD, namely, apoptosis, inflammation, and axon guidance.

Alter et al. examined global expression changes in autism related to paternal age, using peripheral blood lymphocytes from 82 ASD affected children and 64 controls [158]. Cases had a formal autism diagnosis according to DSM-IV, and were confirmed by ADOS and ADI-R, individuals with known genetic disorder were excluded, along with high functioning individuals (Asperger's). Expression profiling was performed using microarray technology designed to measure the expression of 38,500 genes, and standard data processing steps performed. The primary anal-

ysis in the study was of the variance in the distribution of gene expression levels in ASD cases compared to controls, for which the expression values were log2 transformed, and the average squared deviations from the mean calculated. ASD cases were found to have significantly decreased variance in gene expression compared to controls. Taking paternal age into account, significantly decreased overall variance in gene expression was observed for controls in association with increasing paternal age, and the overall variance for children from older fathers was the same as that for children with ASD from fathers of any age. In terms of the specific differences in expression seen, there were many more significantly down-regulated genes in ASD children and children with older fathers. Furthermore, there was a significant overlap between the genes found to be up or down regulated for children with autism and children of older fathers. Pathway analysis of these overlapping genes revealed enrichment for multiple pathways associated with transcriptional regulation, particularly in the down regulated genes. Overall, it was concluded that dysregulation of transcription could underlie the decreased variance in global gene expression in children from older fathers, with increased risk for ASD.

A study by Seno et al. investigated mRNA and miRNA expression in ASD using lymphoblastoid cell lines derived from 18 families comprising 20 severely autistic cases along with their 20 unaffected siblings. Case status was ascertained using the ADOS, and previously generated SNP array data was used to confirm that none of the individuals had detectable CNVs in known susceptibility loci. Gene expression profiling was performed on platforms assaying both mRNA and miRNA, and after following standard data processing steps, differentially expressed genes between cases and controls were identified using Generalized Estimating Equations (GEE). In order to select a final list of genes, it was also required that the genes showed a logFC of at least 1.5 fold in 50% of the sibling pairs. A number of potentially interesting genes were identified including *ARHGAP24*, *IFITM3*, *GBP4*, *HEY1*, and *SOX9*. For six of the identified genes, RT-qPCR was used to confirm the levels of expression. Some miRNAs were also identified as being differentially expressed, the top hit being miR-199b-5p. Pathway analysis subsequently revealed an enrichment for nervous system development and neurological disorders - with Rett syndrome showing the most significant association in this category. Overall,

the authors conclude that their results show that a number of genes and miRNAs important for nervous system development and function are disrupted in ASD, particularly those genes involved in kinase and *NOTCH* signalling. Further, the results suggest a potential overlap with Rett syndrome.

Kuwano et al. examined expression in peripheral blood samples from 21 adolescents and adults with ASD, age-matched controls, and additionally healthy mothers of children with ASD and control mothers with typically developing children. The cases had a pre-existing diagnosis of ASD, which was also confirmed using the DSM-IV. Microarrays were used to assess gene expression across approximately 20,000 genes, and data was processed using the standard protocol. Differential expression analysis was performed using unpaired t test to compare ASD cases with their age-matched controls, with significance thresholds specified at FDR < 0.05 and logFC > 1. A number of genes were found to be up-regulated in cases compared to controls, including *C12ORF58, ITGA2B, LHB, MYOG, NOVA2, NUMBL, PLCXD2, SDK2, TAOK2, UTS2, WDTC1*. Pathway analysis revealed these to be involved in cell morphology, cellular assembly and organization, nerve system development and function. Next, mothers of autistic children were compared to controls using the same analysis strategy as before. Genes coding for ribosomal protein were identified, as well as those involved in immune functions like antigen presentation and cell-mediated immune response. Finally, the results from ASD cases and mothers of ASD cases were compared, which revealed a large degree of overlap in the genes and identified and the direction of their differential expression. Significant results were confirmed by qPCR, which was also used to measure the levels of known ASD candidate genes, where it was found that *FRM1* levels were reduced in ASD cases, and *MECP2* levels were elevated in both ASD and mothers of ASD children. Concluding, the authors state that the genes identified might help reveal the pathology of ASD, with nervous system development and ribosomal pathways identified, along with decreased expression of known ASD candidates.

Voineagu et al. used microarrays to profile gene expression in post-mortem tissue samples from 19 ASD cases and 17 controls, specifically examining the superior temporal gyrus, prefrontal cortex and cerebellar vermis regions [159]. Following data generation and processing following standard protocols, differential expres-

sion analysis was carried out using SAM with significance set at FDR < 0.05 and fold change > 1.3. In total, 444 genes were significantly differentially expressed in ASD compared to controls in the cortex, and there was a significant overlap between identified genes in the frontal and temporal cortex. Only 2 genes were found to be differentially expressed in the cerebellum, which was then excluded from further analysis. RT-PCR was used to confirm the changes for a selection of genes, for which 83% were subsequently confirmed. Gene enrichment analysis showed down regulated genes were enriched for synaptic function, whereas up regulated genes were enriched for genes involved in immune system and inflammatory response. A number of these genes were successfully replicated in another cohort of 9 ASD cases and controls, with many showing expression changes in the same direction as the original discovery cohort, including the *DLX1* and *AHI1*. Higher level, co-expression analysis revealed that regional differences in expression were attenuated in the frontal and temporal lobe in ASD cases. Further co-expression type analysis revealed an overrepresentation of ASD susceptibility genes in the top module of co-expressed genes including *CADSP2, AHI1, CNTNAP2, SLC25A12*. The second module was enriched for astrocyte markers, markers of activated microglia, immune and inflammatory genes. Next, RNA-seq was used to assess differential splicing, where it was confirmed that a number of alternative splicing events are associated of *A2BP1* are associated with ASD. Finally, to help determine the etiology of the observed changes, the co-expression modules were tested for enrichment of previously identified GWAS susceptibility loci, where it was found while the synaptic module was enriched, the immune module was not. Overall, the results were taken as support for the role of synaptic dysfunction and altered immune response in ASD, and further, that synaptic changes are likely to be genetic whereas immune changes are likely to be secondary or caused by environmental factors.

A study by Chow et al. investigated the relationship between altered gene expression in ASD and age using post-mortem dorsolateral prefrontal cortex samples from 9 ASD and 7 control children, and 6 ASD and 11 control adults. Microarrays were used to assay gene expression for 18,626 genes, and data was preprocessed following the standard analytical procedures. To assess differential expression between young autistic, young control, adult autistic, and adult control,

two-way ANOVA was used, and to identify genes showing an effect for interaction between age and diagnosis, an overall pairwise ANOVA-based F-test was used with significance set at FDR < 0.27. In total, 102 genes were identified that showed differential expression in young ASD cases compared with controls and showing an age group interaction effect. Pathway enrichment suggested enrichment for DNA-damage response, cell cycle and apoptosis, immune signalling, neurogenesis and neural development. A number of genes of interest were highlighted including *BRCA1, CHK2, FAS, BCL3, GREM1, FOSL2, FGF1, HOXD1, NDE1, NODAL, PCSK6*. As for the gene expression differences in adult ASD cases compared to controls, 736 genes were identified, which were enriched for cell differentiation, mitogenic signalling and apoptosis. Notable genes included *MAPK12, CDKN1A, NTRK3, PRKAR1A, PIK3CA, CASP9, MAPK10, ADCY6, MAGED1*. Finally, comparing differentially expressed genes between ASD and control independent of age, over 2000 genes were identified. These were enriched for DNA-damage response, apoptosis, and immune system response, and p53-PTEN signalling. Concluding, the results showed that in young autistic brains genes and pathways important for development are disrupted, whereas in adult brains, those important for neurogenesis and repair are disrupted. As for age-independent differences, the results suggested the involvement of immune system response, DNA-damage response, and apoptosis.

Glatt et al. performed expression profiling of peripheral blood mononuclear cells from 60 infants and toddlers at risk for ASDs, 34 at risk for language delay, 17 at risk for global developmental delay, and 68 typically developing controls. Children were initially diagnosed using the ADOS and ADI-R tests. The samples were run on microarrays and the generated data was pre-processed following standard protocols. For the analysis, the sample was split into discovery and replication samples, differentially expressed genes identified using analysis of covariance (ANCOVA) with threshold for significance specified as nominal $p < 0.05$ and logFC $\geq$ 1.2. The identified genes were used to build a support vector machine (SVM) classifier, which was subsequently tested for classification accuracy in the independent replication samples. The results showed 154 genes with significant differential expression in ASD cases compared to controls, which were then used to build, opti-

mize and test the SVM classifier. The final classifier chosen used 48 genes correctly classified 71% of ASD subjects across 10 subsets of the discovery sample into their correct diagnostic categories. When applied to the replication sample, the classifier placed 91% of the subjects into their correct category, and the overall sensitivity was 0.93 and specificity 0.88. Pathway analysis revealed that the 48 genes making up the classifier were enriched for immune response and haemoglobin complex. Concluding, the results provide robust evidence for the existence of a gene expression profile for ASD that is detectable in blood, that could serve as an accurate diagnostic test.

Finally, Kong et al. [160] investigated gene expression in 20 ASD and unaffected sibling pairs and 18 unrelated controls taken from the Simons Simplex Collection (SSC). Diagnoses were made using the ADOS and ADI-R. Gene expression profiling was performed using microarrays measuring a total of 28,869 genes, and the data processed following the standard analysis strategy. In order to identify differentially expressed genes between ASD cases and unaffected sibling, linear models were used, with significance threshold specified at $p < 0.01$. This identified 163 genes which included two previously reported ASD candidates *CTNNB1* and *XPO1*. The subjects were then clustered based on their gene expression profiles, which revealed subgroups of siblings, displaying higher similarity to ASD cases - termed proband-like, and less similar – termed control-like. Geneset enrichment analysis revealed that ribosome and spliceosomal pathways were up regulated and immune signalling downregulated in ASD cases compared to proband-like siblings. Concluding, the gene expression profiles generated were able to distinguish ASD affected from unaffected siblings, and further implicated upregulation of ribosomal and mitochondrial genes, and down regulation of immune response pathways.

The overall pattern emerging from gene expression studies is of disruption of normal transcriptional control, and the potential involvement of an immune or inflammatory component in ASD. This perhaps in contrast to genetic studies which have tended to implicate synaptic organization and inhibitory/excitatory pathways, therefore the identification of these alternative pathways to ASD suggest that gene expression profiling could be revealing another dimension to ASD pathology. A

less generous interpretation might be that the frequent identification of these pathways in fact reflects limitations with the genome-wide expression approach, particularly the use of non-primary tissues. This would seem unlikely however, as the findings seem to be consistent across the majority of the studies, which use different cohorts, populations, diagnostic criteria, and tissue types - including cortex and cerebellar samples. Some limitations of these previous studies are that firstly, with the exception of the Hu et al. study [152], they are not designed to separate gene expression differences resulting from non-shared environmental factors, to those arising from shared environment or genetic factors. Secondly, due to the limitations of microarray profiling technology, they investigate only transcript abundance, and by doing so could be missing alternative splicing and allele-specific expression events associated with ASD. Newer techniques such as RNA-seq enable these other transcriptomic phenomena to be investigated.

1.7 Gene expression profiling of monozygotic twins with ASD

1.7.1 Monozygotic twins

There are many challenges associated with transcriptomic and epigenetic investigations, especially those comparing affected cases with unaffected controls. Because unrelated individuals possess distinct genetic backgrounds, any observed differences at the level of gene expression or DNA methylation can be genetic or non-genetic in origin. Using population-base case-control studies, it is often not possible to disentangle the relative contribution of these or to determine if the differences are likely to be driven by environmental factors. Primarily for these reasons, some authors have suggested utilising twin-based study designs more generally in the study of complex traits and conditions [161, 162, 163, 164, 165].

MZ twins are genetically identical in the absence of any identified rare, non-inherited genetic variation, yet can often display divergent phenotypes. This has been ob-

served for a range of neurodevelopmental conditions including schizophrenia [166], bipolar disorder [167], and ASD [25]. Incomplete concordance in MZ twins suggests a causal role for epigenetic, stochastic, and/or environmental factors. In support of this, variations in gene expression and DNA methylation are observed in MZ twins, which may in turn underlie phenotypic variation [105, 163]. A study design comparing discordant twins enables the molecular basis of trait discordance to be explored, as well as providing a means to separate genetic from non-genetic influence. This is possible because MZ twins, in addition to being genetically identical, are also matched for age, gender, maternal environment, population cohort effects, and exposure to other shared environmental factors [161]. Any identified phenotypic differences (and by extension, molecular differences) are then thought to be predominantly attributable to non-genetic influences [168, 169].

1.7.2 Gene expression profiling of MZ twins with ASD

In Chapters 2 and 3, we describe two gene expression studies making up a large part of the empirical work undertaken for this thesis. We utilise a twins-based design with ASD concordant, ASD discordant, and matched control MZ pairs for whom whole-blood samples and behavioural measurements are available. Gene expression profiles are generated using both microarray and RNA-seq technologies, with the overall aim to investigate patterns of gene expression in ASD, and identify the genes and pathways disrupted. Further, by focusing on differences within the MZ discordant pairs, we hope to identify which perturbations are responsive to the non-shared environment.

1.7.3 Potential limitations of this study

There are a number of potential limitations associated with the proposed methods. As we will later rely on the data produced by our profiling experiments for integrative analysis (introduced in the next section), it is worth spending some time now considering each of these limitations in turn, as they ultimately have implications for the types of inferences that can be made.

1.7.3.1 Are gene expression measurements reliable?

Genome-wide approaches are predicated on the robustness of genome-wide gene expression signatures. This is almost taken as axiomatic, given that transcriptomics methods are so well established, but nonetheless, a large meta-analysis by Dudley et al. showed expression profiles to be highly concordant between disease states, across tissues, and even across separate studies [170].

1.7.3.2 Do mRNA levels correspond to protein levels?

The question of to what extent mRNA levels actually correspond to protein levels, is an important one, but one that is not easily answered. Post-transcriptional mechanisms in the cell, such as RNAi by ncRNAs, can regulate levels of mRNA, meaning caution must be used when attempting to interpret the functional implications of increased mRNA expression. As a general rule, increased mRNA expression can be taken as indicative of an increased level of the protein product, however this relationship should ideally be determined on a gene-by-gene basis and in the context of the trait being studied - to investigate potential quantitative or gene dosage effects. Previous expression studies into ASD have for example, quantified the abundance of protein products for genes previously identified as differentially expressed (see earlier reviews of the Purcell et al. [151] and Hu et al. papers [157]). We may choose a similar approach in a follow-up study, should we identify any coding genes showing evidence of differential expression in our ASD twins.

1.7.3.3 Are expression differences between MZ twins detectable?

An important question in the context of this project, is whether expression differences between MZ twins are large enough to be detectable. A number of studies have investigated this more generally. Firstly, Sharma et al. assessed natural variation in gene expression for 5 pairs of MZ twins and compared this to 13 unrelated individuals using microarrays [105]. Here it was found that < 2% of genes showed differential expression between twins, whereas for unrelated individuals this was

much higher, at 14%. On the basis of the results it was then suggested expression profiling of MZ twins could be valuable for identifying environmentally sensitive expression loci. Another study by Cheung et al. examined natural variation in gene expression in lymphoblastoid cells for a sample of 35 individuals and for 10 pairs of MZ twins with microarray[171]. For a number of genes that were identified as highly variable between unrelated individuals, the variance in expression levels between MZ twins was found to display only $1/3$ to $1/11$ variation. The results were taken to indicate that a component of variation in gene expression is likely genetically determined, but the authors did not make any statement about the likely nature of non-genetic component.

Previous estimates for sample size requirements for gene expression profiling have shown that around 10 cases and 10 controls would be required to detect gene expression changes of 2 SDs with 80% power at an FDR of 5% [172, 173, 174]. These are based on case control studies, and though we were unable to find any calculation that estimated this for a twin-based study design, it seems reasonable to postulate that an MZ twins study would have greater power due to twins being matched for genotype and certain shared environmental influences. Indeed, support for this comes from previous work by Vischer and Posthuma, where power to detect phenotypic variance due to environmental effects was calculated for both studies utilising unrelated subjects and MZ-twins [175]. Here, it was found that for traits with MZ correlation > 0.3, the MZ design is more efficient than one using unrelated individuals. There have also been a number of studies that have looked at power and sample sizes required for twin-based studies in the context of EWAS. Kaminsky et al. estimated that 15-25 discordant twin pairs would be sufficient for 80% power to detect DNA methylation differences of 1.2 fold change [176]. Bell and Tsai used simulations to quantify the sample sizes required to reach 80% power in EWAS over a range of different effect sizes for both case-control and MZ-differences study designs [177]. They reported a minimum required sample size of 30 pairs of twins compared to 37 cases in order to detect nominally significant mean methylation differences of 7% using a parametric test. Indeed, smaller sample sizes were estimated for twins as compared to unrelated individuals over a range of different effect sizes from 7% to 15%. In sum, taking into account the previous

power calculations for environmental variance and sample size estimations for detecting DNA methylation differences between MZ twins, the expectation for this study is that we should similarly be able to detect transcriptomic differences in our sample of MZ twin pairs.

1.7.3.4 Whole blood as a surrogate

Ideally, we would profile disease associated transcriptomic differences in the primary affected tissue, but in neurodevelopmental disorders due to the inaccessibility of brain tissue, often peripheral tissues are instead used. Whole blood is typically seen as a reasonable surrogate because it is easily accessible and collection is minimally invasive, and since it circulates throughout the entire body coming into contact with all other tissues. The rationale for using whole blood is then that it might be possible to detect disease-associated changes in expression in other tissues at low levels, that more widespread global patterns of disruption might be revealed, or finally, that blood cells could themselves respond to the same insult that lead to the expression changes in the primary affected tissue - acting as a "sentinel" for disease [178].

With neurodevelopmental disorders there is the added complication of the blood brain barrier - which could prevent brain-specific transcripts from diffusing into blood altogether, as well as the temporal aspect of the disorder - where pathogenesis likely occured during development. In this case, the utility of blood as a surrogate is perhaps less certain. The first step to addressing this question is to determine the overall concordance between genes expressed in brain and those in blood, and in fact several studies have attempted to do so. Liew et al. compared the genes expressed in whole blood with those from nine other tissues including brain, and found that between 66% to 82% of all coding genes were detectable in blood, and that 81.9% of genes expressed in brain were also expressed in blood (based on the overlap of identified genes) [178]. A more recent study by Cai et al. compared expression profiles generated for cortex, cerebellum, and caudate nucleus, to blood and looked at correlation between mean expression levels, as well as the preservation of brain gene co-expression modules in blood [179]. The overall level of cor-

relation of expression between blood and brain was found to be weak, in the range of 0.24 to 0.32, and while only a small number of brain co-expression modules were found to be preserved in blood, those that were showed strong preservation, suggesting their potential utility as biomarkers in neurodevelopmental investigations.

One major limitation of looking at overall concordance between mean expression levels in brain and blood is that it does not really address the main issue which is whether any differences in expression are detectable. While this is difficult to address in human tissue, a study by Davies et al. investigated this in rats using a panel of recombinant inbred strains for hippocampus and spleen [180]. Overall it was found that a large number of genes were expressed in common between the two different types of tissue, but even more pertinently, that the correlation between expression depends on the variance of the transcripts across lines, with those that vary more (i.e. showing higher heritability of expression) more highly correlated and hence are more likely to be detectable in the surrogate. From this, we conclude that surrogate tissues are likely to be useful for detect genetically driven expression differences, but their utility for detecting non-genetically driven expression differences, as we are attempting to identify in our MZ ASD twins, perhaps then remains to be demonstrated.

Finally, one last important consideration in using blood as a surrogate is whether individual differences in gene expression as assessed are stable over time. Meaburn et al. examined the reliability of detected individual differences in gene expression in blood samples, and found that these differences could be reliably detected 10 months after collection[108].

1.8 Integrating multiple genomics datasets

Increasingly it is appreciated that in order to make further progress in uncovering the etiology and mechanisms underlying ASD as well as other complex traits, a more holistic approach will be required to bring together evidence from many disparate experimental sources and attempt to understand the dynamic workings of

the system as a whole [129]. Within ASD genomics research, this will mean combining data from the multiple molecular modalities in order to really pick apart the relationship between genetic variation, gene regulatory mechanisms, environmental perturbation of these regulatory mechanisms, typical brain development, and autistic traits. The need for a system-level perspective on ASD is rather neatly illustrated by the observation that while the accumulated findings from genetic association studies suggest that ASD is primarily a disorder of the synapse, functional approaches like expression and methylomic profiling implicate immune dysregulation, inflammation, and epigenetic/transcriptomic disruption. To this end, integrative genomics approaches aim to combine evidence from different layers of the genome in order to boost power to identify molecular associations and provide a more coherent picture of the pathomechanisms underlying complex diseases. As we shall later discuss, quality control is an important step in data integration, and so in Chapter 4 we address an issue related to EWAS - namely the empirical calculation of an appropriate significance threshold for single CpG site differential methylation analysis. In Chapter 5, we then delve deeper into integrative methods, applying some of these to the gene expression data generated in Chapter 3 and combining with an existing methylation dataset on the same sample.

References

[1] L. Kanner *et al.*, "Autistic disturbances of affective contact," 1943.

[2] L. Kanner, "Irrelevant and metaphorical language in early infantile autism," *American journal of Psychiatry*, vol. 103, no. 2, pp. 242–246, 1946.

[3] A. P. Association *et al.*, *DSM 5*. American Psychiatric Association, 2013.

[4] A. P. Association *et al.*, "Diagnostic and statistical manual of mental disorders dsm-iv-tr fourth edition (text revision)," 2000.

[5] G. Baird, E. Simonoff, A. Pickles, S. Chandler, T. Loucas, D. Meldrum, and T. Charman, "Prevalence of disorders of the autism spectrum in a population cohort of children in south thames: the special needs and autism project (snap)," *The lancet*, vol. 368, no. 9531, pp. 210–215, 2006.

[6] M. Butler, M. Dasouki, X. Zhou, Z. Talebizadeh, M. Brown, T. Takahashi, J. Miles, C. Wang, R. Stratton, R. Pilarski, *et al.*, "Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline pten tumour suppressor gene mutations," *Journal of medical genetics*, vol. 42, no. 4, pp. 318–321, 2005.

[7] R. Tuchman and I. Rapin, "Epilepsy in autism," *The Lancet Neurology*, vol. 1, no. 6, pp. 352–358, 2002.

[8] E. Shishido, B. Aleksic, and N. Ozaki, "Copy-number variation in the pathogenesis of autism spectrum disorder," *Psychiatry and clinical neurosciences*, vol. 68, no. 2, pp. 85–95, 2014.

[9] J. Chen, I. Alberts, and X. Li, "Dysregulation of the igf-i/pi3k/akt/mtor signaling pathway in autism spectrum disorders," *International Journal of Developmental Neuroscience*, vol. 35, pp. 35–41, 2014.

[10] V. Lotter, "Epidemiology of autistic conditions in young children," *Social psychiatry*, vol. 1, no. 3, pp. 124–135, 1966.

[11] M. Parellada, M. Penzol, L. Pina, C. Moreno, E. González-Vioque, G. Zalsman, and C. Arango, "The neurobiology of autism spectrum disorders," *European Psychiatry*, vol. 29, no. 1, pp. 11–19, 2014.

[12] M. Elsabbagh, G. Divan, Y.-J. Koh, Y. S. Kim, S. Kauchali, C. Marcín, C. Montiel-Nava, V. Patel, C. S. Paula, C. Wang, *et al.*, "Global prevalence of autism and other pervasive developmental disorders," *Autism Research*, vol. 5, no. 3, pp. 160–179, 2012.

[13] T. Brugha, S. McManus, H. Meltzer, J. Smith, F. Scott, S. Purdon, J. Harris, and J. Bankart, "Autism spectrum disorders in adults living in households throughout england: Report from the adult psychiatric morbidity survey 2007," *Leeds: The NHS Information Centre for Health and Social Care*, 2009.

[14] G. Russell, "The rise and rise of the autism diagnosis," *Autism-Open Access*, vol. 2012, 2012.

[15] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, *et al.*, "Rate of de novo mutations and the importance of father/'s age to disease risk," *Nature*, vol. 488, no. 7412, pp. 471–475, 2012.

[16] P. Szatmari, M. Jones, L. Zwaigenbaum, and J. MacLean, "Genetics of autism: overview and new directions," *Journal of autism and developmental disorders*, vol. 28, no. 5, pp. 351–368, 1998.

[17] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Dobkins, *et al.*, "Recurrence risk for autism spectrum disorders: a baby siblings research consortium study," *Pediatrics*, vol. 128, no. 3, pp. e488–e495, 2011.

[18] J. N. Constantino, C. Lajonchere, M. Lutz, T. Gray, A. Abbacchi, K. McKenna, D. Singh, and R. D. Todd, "Autistic social impairment in the siblings of children with pervasive developmental disorders," *American Journal of Psychiatry*, 2006.

[19] M. Losh, R. Adolphs, M. D. Poe, S. Couture, D. Penn, G. T. Baranek, and J. Piven, "Neuropsychological profile of autism and the broad autism phenotype," *Archives of general psychiatry*, vol. 66, no. 5, pp. 518–526, 2009.

[20] S. Folstein and M. Rutter, "Infantile autism: a genetic study of 21 twin pairs," *Journal of Child psychology and Psychiatry*, vol. 18, no. 4, pp. 297–321, 1977.

[21] S. Steffenburg, C. Gillberg, L. Hellgren, L. Andersson, I. C. Gillberg, G. Jakobsson, and M. Bohman, "A twin study of autism in denmark, finland, iceland, norway and sweden," *Journal of Child Psychology and Psychiatry*, vol. 30, no. 3, pp. 405–416, 1989.

[22] A. Bailey, A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda, and M. Rutter, "Autism as a strongly genetic disorder: evidence from a british twin study," *Psychological medicine*, vol. 25, no. 01, pp. 63–77, 1995.

[23] R. E. Rosenberg, J. K. Law, G. Yenokyan, J. McGready, W. E. Kaufmann, and P. A. Law, "Characteristics and concordance of autism spectrum disorders among 277 twin pairs," *Archives of pediatrics & adolescent medicine*, vol. 163, no. 10, pp. 907–914, 2009.

[24] J. Hallmayer, S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, *et al.*, "Genetic heritability and shared environmental factors among twin pairs with autism," *Archives of general psychiatry*, vol. 68, no. 11, p. 1095, 2011.

[25] A. Ronald and R. A. Hoekstra, "Autism spectrum disorders and autistic traits: a decade of new twin studies," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 156, no. 3, pp. 255–274, 2011.

[26] B. Devlin and S. W. Scherer, "Genetic architecture in autism spectrum disorder," *Current opinion in genetics & development*, vol. 22, no. 3, pp. 229–237,

2012.

[27] D. S. Falconer, *Introduction to quantitative genetics*. Pearson Education India, 1975.

[28] S. Tordjman, E. Somogyi, N. Coulon, S. Kermarrec, D. Cohen, G. Bronsard, O. Bonnot, C. Weismann-Arcache, M. Botbol, B. Lauth, *et al.*, "Gene× environment interactions in autism spectrum disorders: role of epigenetic mechanisms," *Frontiers in psychiatry*, vol. 5, p. 53, 2014.

[29] C. Betancur, "Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting," *Brain research*, vol. 1380, pp. 42–77, 2011.

[30] J. Vorstman, W. Staal, E. Van Daalen, H. Van Engeland, P. Hochstenbach, and L. Franke, "Identification of novel autism candidate regions through analysis of reported cytogenetic abnormalities associated with autism," *Molecular psychiatry*, vol. 11, no. 1, pp. 18–28, 2006.

[31] G. Huguet, E. Ey, and T. Bourgeron, "The genetic landscapes of autism spectrum disorders," *Annual review of genomics and human genetics*, vol. 14, pp. 191–213, 2013.

[32] S. E. Folstein and B. Rosen-Sheidley, "Genetics of austim: complex aetiology for a heterogeneous disorder," *Nature Reviews Genetics*, vol. 2, no. 12, pp. 943–955, 2001.

[33] C. M. Freitag, "The genetics of autistic disorders and its clinical relevance: a review of the literature," *Molecular psychiatry*, vol. 12, no. 1, pp. 2–22, 2007.

[34] D. H. Geschwind, "Genetics of autism spectrum disorders," *Trends in cognitive sciences*, vol. 15, no. 9, pp. 409–416, 2011.

[35] V. Brown, P. Jin, S. Ceman, J. C. Darnell, W. T. O'Donnell, S. A. Tenenbaum, X. Jin, Y. Feng, K. D. Wilkinson, J. D. Keene, *et al.*, "Microarray identification of fmrp-associated brain mrnas and altered mrna translational profiles in fragile x syndrome," *Cell*, vol. 107, no. 4, pp. 477–487, 2001.

[36] T. Kishino, M. Lalande, and J. Wagstaff, "Ube3a/e6-ap mutations cause angelman syndrome," *Nature genetics*, vol. 15, no. 1, pp. 70–73, 1997.

[37] S. Jamain, H. Quach, C. Betancur, M. Råstam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, *et al.*, "Mutations of the x-linked genes encoding neuroligins nlgn3 and nlgn4 are associated with autism," *Nature genetics*, vol. 34, no. 1, pp. 27–29, 2003.

[38] H.-G. Kim, S. Kishikawa, A. W. Higgins, I.-S. Seong, D. J. Donovan, Y. Shen, E. Lally, L. A. Weiss, J. Najm, K. Kutsche, *et al.*, "Disruption of neurexin 1 associated with autism spectrum disorder," *The American Journal of Human Genetics*, vol. 82, no. 1, pp. 199–207, 2008.

[39] S. Berkel, C. R. Marshall, B. Weiss, J. Howe, R. Roeth, U. Moog, V. Endris, W. Roberts, P. Szatmari, D. Pinto, *et al.*, "Mutations in the shank2 synaptic scaffolding gene in autism spectrum disorder and mental retardation," *Nature genetics*, vol. 42, no. 6, pp. 489–491, 2010.

[40] C. M. Durand, C. Betancur, T. M. Boeckers, J. Bockmann, P. Chaste, F. Fauchereau, G. Nygren, M. Rastam, I. C. Gillberg, H. Anckarsäter, *et al.*, "Mutations in the gene encoding the synaptic scaffolding protein shank3 are associated with autism spectrum disorders," *Nature genetics*, vol. 39, no. 1, pp. 25–27, 2007.

[41] A. Zuko, K. T. Kleijer, A. Oguro-Ando, M. J. Kas, E. van Daalen, B. van der Zwaag, and J. P. H. Burbach, "Contactins in the neurobiology of autism," *European journal of pharmacology*, vol. 719, no. 1, pp. 63–74, 2013.

[42] D. E. Arking, D. J. Cutler, C. W. Brune, T. M. Teslovich, K. West, M. Ikeda, A. Rea, M. Guy, S. Lin, E. H. Cook, *et al.*, "A common genetic variant in the neurexin superfamily member cntnap2 increases familial risk of autism," *The American Journal of Human Genetics*, vol. 82, no. 1, pp. 160–164, 2008.

[43] D. Malhotra and J. Sebat, "Cnvs: harbingers of a rare variant revolution in psychiatric genetics," *Cell*, vol. 148, no. 6, pp. 1223–1241, 2012.

[44] P. Szatmari, A. D. Paterson, L. Zwaigenbaum, W. Roberts, J. Brian, X.-Q. Liu, J. B. Vincent, J. L. Skaug, A. P. Thompson, L. Senman, *et al.*, "Mapping autism risk loci using genetic linkage and chromosomal rearrangements," *Nature genetics*, vol. 39, no. 3, pp. 319–328, 2007.

[45] B. H.-Y. Chung, V. Q. Tao, and W. W.-Y. Tso, "Copy number variation and autism: new insights and clinical implications," *Journal of the Formosan Medical Association*, vol. 113, no. 7, pp. 400–408, 2014.

[46] J. T. Glessner, K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, *et al.*, "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes," *Nature*, vol. 459, no. 7246, pp. 569–573, 2009.

[47] D. Sato, A. C. Lionel, C. S. Leblond, A. Prasad, D. Pinto, S. Walker, I. O'Connor, C. Russell, I. E. Drmic, F. F. Hamdan, *et al.*, "Shank1 deletions in males with autism spectrum disorder," *The American Journal of Human Genetics*, vol. 90, no. 5, pp. 879–887, 2012.

[48] A. T. Pagnamenta, H. Khan, S. Walker, D. Gerrelli, K. Wing, M. C. Bonaglia, R. Giorda, T. Berney, E. Mani, M. Molteni, *et al.*, "Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (cdh8) in susceptibility to autism and learning disability," *Journal of medical genetics*, vol. 48, no. 1, pp. 48–54, 2011.

[49] L. A. Weiss, D. E. Arking, M. J. Daly, A. Chakravarti, C. W. Brune, K. West, A. O'Connor, G. Hilton, R. L. Tomlinson, A. B. West, *et al.*, "A genome-wide linkage and association scan reveals novel loci for autism," *Nature*, vol. 461, no. 7265, pp. 802–808, 2009.

[50] R. Anney, L. Klei, D. Pinto, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, N. Sykes, A. T. Pagnamenta, *et al.*, "A genome-wide scan for common alleles affecting risk for autism," *Human molecular genetics*, p. ddq307, 2010.

[51] K. Wang, H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, *et al.*, "Common genetic variants on 5p14. 1 associate with autism spectrum disorders," *Nature*, vol. 459, no. 7246, pp. 528–533, 2009.

[52] B. Torrico, A. G. Chiocchetti, E. Bacchelli, E. Trabetti, A. Hervás, B. Franke, J. K. Buitelaar, N. Rommelse, A. Yousaf, E. Duketis, *et al.*, "Lack of replication of previous autism spectrum disorder gwas hits in european populations," *Autism Research*, 2016.

[53] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *TRENDS in Genetics*, vol. 17, no. 9, pp. 502–510, 2001.

[54] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nature genetics*, vol. 40, no. 6, pp. 695–701, 2008.

[55] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[56] M. E. Goddard, N. R. Wray, K. Verbyla, P. M. Visscher, *et al.*, "Estimating effects and making predictions from genome-wide marker data," *Statistical Science*, vol. 24, no. 4, pp. 517–529, 2009.

[57] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, *et al.*, "Common snps explain a large proportion of the heritability for human height," *Nature genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[58] M. Manchia, J. Cullis, G. Turecki, G. A. Rouleau, R. Uher, and M. Alda, "The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases," *PLoS One*, vol. 8, no. 10, p. e76295, 2013.

[59] L. Klei, S. J. Sanders, M. T. Murtha, V. Hus, J. K. Lowe, A. J. Willsey, D. Moreno-De-Luca, W. Y. Timothy, E. Fombonne, D. Geschwind, *et al.*, "Common genetic variants, acting additively, are a major source of risk for autism," *Molecular autism*, vol. 3, no. 1, p. 1, 2012.

[60] C.-D. G. of the Psychiatric Genomics Consortium *et al.*, "Genetic relationship between five psychiatric disorders estimated from genome-wide snps," *Nature genetics*, vol. 45, no. 9, pp. 984–994, 2013.

[61] T. Gaugler, L. Klei, S. J. Sanders, C. A. Bodea, A. P. Goldberg, A. B. Lee, M. Mahajan, D. Manaa, Y. Pawitan, J. Reichert, *et al.*, "Most genetic risk for autism resides with common variation," *Nature genetics*, vol. 46, no. 8, pp. 881–885, 2014.

[62] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "Gcta: a tool for genome-wide complex trait analysis," *The American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011.

[63] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability," *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1193–1198, 2012.

[64] E. M. Derks, C. V. Dolan, and D. I. Boomsma, "A test of the equal environment assumption (eea) in multivariate twin studies," *Twin Research and Human Genetics*, vol. 9, no. 03, pp. 403–411, 2006.

[65] C. R. Marshall, A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, *et al.*, "Structural variation of chromosomes in autism spectrum disorder," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 477–488, 2008.

[66] N. Krumm, B. J. O'Roak, E. Karakoc, K. Mohajeri, B. Nelson, L. Vives, S. Jacquemont, J. Munson, R. Bernier, and E. E. Eichler, "Transmission disequilibrium of small cnvs in simplex autism," *The American Journal of Human Genetics*, vol. 93, no. 4, pp. 595–606, 2013.

[67] D. Levy, M. Ronemus, B. Yamrom, Y.-h. Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, *et al.*, "Rare de novo and transmitted copy-number variation in autistic spectrum disorders," *Neuron*, vol. 70, no. 5, pp. 886–897, 2011.

[68] S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, *et al.*, "Multiple recurrent de novo cnvs, including duplications of the 7q11. 23 williams syndrome region, are strongly associated with autism," *Neuron*, vol. 70, no. 5, pp. 863–885, 2011.

[69] I. Iossifov, M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum, B. Yamrom, Y.-h. Lee, G. Narzisi, A. Leotta, *et al.*, "De novo gene disruptions in children on the autistic spectrum," *Neuron*, vol. 74, no. 2, pp. 285–299, 2012.

[70] S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, *et al.*, "De novo mutations revealed by whole-exome sequencing are strongly associated with autism," *Nature*, vol. 485, no. 7397, pp. 237–241, 2012.

[71] Z. A. Kaminsky, T. Tang, S.-C. Wang, C. Ptak, G. H. Oh, A. H. Wong, L. A. Feldcamp, C. Virtanen, J. Halfvarson, C. Tysk, *et al.*, "Dna methylation profiles in monozygotic and dizygotic twins," *Nature genetics*, vol. 41, no. 2, pp. 240–245, 2009.

[72] A. Tenesa and C. S. Haley, "The heritability of human disease: estimation, uses and abuses," *Nature Reviews Genetics*, vol. 14, no. 2, pp. 139–149, 2013.

[73] J. N. Constantino and R. D. Todd, "Genetic structure of reciprocal social behavior," *American Journal of Psychiatry*, vol. 157, no. 12, pp. 2043–2045, 2000.

[74] J. N. Constantino and R. D. Todd, "Autistic traits in the general population: a twin study," *Archives of general psychiatry*, vol. 60, no. 5, pp. 524–530, 2003.

[75] A. Ronald, F. Happé, and R. Plomin, "The genetic relationship between individual differences in social and nonsocial behaviours characteristic of autism," *Developmental science*, vol. 8, no. 5, pp. 444–458, 2005.

[76] D. H. Skuse, W. P. Mandy, and J. Scourfield, "Measuring autistic traits: heritability, reliability and validity of the social and communication disorders checklist," *The British Journal of Psychiatry*, vol. 187, no. 6, pp. 568–572, 2005.

[77] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, no. 4, pp. 669–681, 2007.

[78] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002.

[79] A. Wutz and R. Jaenisch, "A shift from reversible to irreversible x inactivation is triggered during es cell differentiation," *Molecular cell*, vol. 5, no. 4, pp. 695–705, 2000.

[80] J. T. Lee and M. S. Bartolomei, "X-inactivation, imprinting, and long noncoding rnas in health and disease," *Cell*, vol. 152, no. 6, pp. 1308–1323, 2013.

[81] A. C. Ferguson-Smith and M. A. Surani, "Imprinting and the epigenetic asymmetry between parental genomes," *Science*, vol. 293, no. 5532, pp. 1086–1089, 2001.

[82] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[83] J. Mill and B. T. Heijmans, "From promises to practical strategies in epigenetic epidemiology," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 585–594, 2013.

[84] A. M. Deaton and A. Bird, "Cpg islands and the regulation of transcription," *Genes & development*, vol. 25, no. 10, pp. 1010–1022, 2011.

[85] P. A. Jones, "The dna methylation paradox," *Trends in Genetics*, vol. 15, no. 1, pp. 34–37, 1999.

[86] S. Feng, S. E. Jacobsen, and W. Reik, "Epigenetic reprogramming in plant and animal development," *Science*, vol. 330, no. 6004, pp. 622–627, 2010.

[87] S. A. Smallwood and G. Kelsey, "De novo dna methylation: a germ cell perspective," *Trends in Genetics*, vol. 28, no. 1, pp. 33–42, 2012.

[88] C. Popp, W. Dean, S. Feng, S. J. Cokus, S. Andrews, M. Pellegrini, S. E. Jacobsen, and W. Reik, "Genome-wide erasure of dna methylation in mouse primordial germ cells is affected by aid deficiency," *Nature*, vol. 463, no. 7284, pp. 1101–1105, 2010.

[89] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 5, pp. 1827–1831, 1992.

[90] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller, "Validation of a dna methylation microarray for 450,000 cpg sites in the human genome," *Epigenetics*, vol. 6, no. 6, pp. 692–702, 2011.

[91] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450k technology," *Epigenomics*, vol. 3, no. 6, pp. 771–784, 2011.

[92] M. Bibikova, J. Le, B. Barnes, S. Saedinia-Melnyk, L. Zhou, R. Shen, and K. L. Gunderson, "Genome-wide dna methylation profiling using infinium® assay," *Epigenomics*, vol. 1, no. 1, pp. 177–200, 2009.

[93] D. Barker, "The developmental origins of adult disease," *Journal of the American College of Nutrition*, vol. 23, no. sup6, pp. 588S–595S, 2004.

[94] V. K. Rakyan and S. Beck, "Epigenetic variation and inheritance in mammals," *Current opinion in genetics & development*, vol. 16, no. 6, pp. 573–577, 2006.

[95] H. T. Bjornsson, M. D. Fallin, and A. P. Feinberg, "An integrated epigenetic and genetic approach to common human disease," *TRENDS in Genetics*, vol. 20, no. 8, pp. 350–358, 2004.

[96] S. Kigar and A. Auger, "Epigenetic mechanisms may underlie the aetiology of sex differences in mental health risk and resilience," *Journal of neuroendocrinology*, vol. 25, no. 11, pp. 1141–1150, 2013.

[97] T. Kato, K. Iwamoto, C. Kakiuchi, G. Kuratomi, and Y. Okazaki, "Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders," *Molecular psychiatry*, vol. 10, no. 7, pp. 622–630, 2005.

[98] H. Heyn, F. J. Carmona, A. Gomez, H. J. Ferreira, J. T. Bell, S. Sayols, K. Ward, O. A. Stefansson, S. Moran, J. Sandoval, *et al.*, "Dna methylation profiling in breast cancer discordant identical twins identifies dok7 as novel epigenetic biomarker," *Carcinogenesis*, vol. 34, no. 1, pp. 102–108, 2013.

[99] K. Walter, T. Holcomb, T. Januario, P. Du, M. Evangelista, N. Kartha, L. Iniguez, R. Soriano, L. Huw, H. Stern, *et al.*, "Dna methylation profiling defines clinically relevant biological subsets of non–small cell lung cancer," *Clinical Cancer Research*, vol. 18, no. 8, pp. 2360–2373, 2012.

[100] Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, *et al.*, "Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis," *Nature biotechnology*, vol. 31, no. 2, pp. 142–147, 2013.

[101] E. Swan, A. Maxwell, and A. McKnight, "Distinct methylation patterns in genes that affect mitochondrial function are associated with kidney disease in blood-derived dna from individuals with type 1 diabetes," *Diabetic Medicine*, 2015.

[102] J. I. Feinberg, K. M. Bakulski, A. E. Jaffe, R. Tryggvadottir, S. C. Brown, L. R. Goldman, L. A. Croen, I. Hertz-Picciotto, C. J. Newschaffer, M. D. Fallin,

et al., "Paternal sperm dna methylation associated with early signs of autism risk in an autism-enriched cohort," *International journal of epidemiology*, p. dyv028, 2015.

[103] Y. Song, K. Miyaki, T. Suzuki, Y. Sasaki, A. Tsutsumi, N. Kawakami, A. Shimazu, M. Takahashi, A. Inoue, C. Kan, *et al.*, "Altered dna methylation status of human brain derived neurotrophis factor gene could be useful as biomarker of depression," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 165, no. 4, pp. 357–364, 2014.

[104] E. Walton, J. Hass, J. Liu, J. L. Roffman, F. Bernardoni, V. Roessner, M. Kirsch, G. Schackert, V. Calhoun, and S. Ehrlich, "Correspondence of dna methylation between blood and brain tissue and its application to schizophrenia research," *Schizophrenia bulletin*, p. sbv074, 2015.

[105] A. Sharma, V. K. Sharma, S. Horn-Saban, D. Lancet, S. Ramachandran, and S. K. Brahmachari, "Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays," *Physiological Genomics*, vol. 21, no. 1, pp. 117–123, 2005.

[106] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, *et al.*, "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, vol. 422, no. 6929, pp. 297–302, 2003.

[107] S. Monks, A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak, S. Edwards, J. Phillips, A. Sachs, and E. Schadt, "Genetic inheritance of gene expression in human cell lines," *The American Journal of Human Genetics*, vol. 75, no. 6, pp. 1094–1105, 2004.

[108] E. L. Meaburn, C. Fernandes, I. W. Craig, R. Plomin, and L. C. Schalkwyk, "Assessing individual differences in genome-wide gene expression in human whole blood: reliability over four hours and stability over 10 months," *Twin Research and Human Genetics*, vol. 12, no. 04, pp. 372–380, 2009.

[109] M. Pheasant and J. S. Mattick, "Raising the estimate of functional human sequences," *Genome research*, vol. 17, no. 9, pp. 1245–1253, 2007.

[110] M. Pertea, "The human transcriptome: an unfinished story," *Genes*, vol. 3, no. 3, pp. 344–360, 2012.

[111] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding rnas," *Molecular cell*, vol. 43, no. 6, pp. 904–914, 2011.

[112] D. B. Pontier and J. Gribnau, "Xist regulation and function explored," *Human genetics*, vol. 130, no. 2, pp. 223–236, 2011.

[113] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, *et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology*, vol. 14, no. 13, pp. 1675–1680, 1996.

[114] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," *Nature genetics*, vol. 21, pp. 20–24, 1999.

[115] D. J. Lockhart and E. A. Winzeler, "Genomics, gene expression and dna arrays," *nature*, vol. 405, no. 6788, pp. 827–836, 2000.

[116] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[117] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, pp. 621–628, 2008.

[118] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, *et al.*, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, no. 5891, pp. 956–960, 2008.

[119] M. N. Cabili, C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev, and J. L. Rinn, "Integrative annotation of human large intergenic noncoding

rnas reveals global properties and specific subclasses," *Genes & development*, vol. 25, no. 18, pp. 1915–1927, 2011.

[120] J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev, "Comprehensive comparative analysis of strand-specific rna sequencing methods," *Nature methods*, vol. 7, no. 9, pp. 709–715, 2010.

[121] R. Toro, M. Konyukh, R. Delorme, C. Leblond, P. Chaste, F. Fauchereau, M. Coleman, M. Leboyer, C. Gillberg, and T. Bourgeron, "Key role for gene dosage and synaptic homeostasis in autism spectrum disorders," *Trends in genetics*, vol. 26, no. 8, pp. 363–372, 2010.

[122] N. C. Schanen, "Epigenetics of autism spectrum disorders," *Human molecular genetics*, vol. 15, no. suppl 2, pp. R138–R150, 2006.

[123] Y. Kuwano, Y. Kamio, T. Kawai, S. Katsuura, N. Inada, A. Takaki, and K. Rokutan, "Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children," *PloS one*, vol. 6, no. 9, p. e24723, 2011.

[124] R. M. Carney, C. M. Wolpert, S. A. Ravan, M. Shahbazian, A. Ashley-Koch, M. L. Cuccaro, J. M. Vance, and M. A. Pericak-Vance, "Identification of mecp2 mutations in a series of females with autistic disorder," *Pediatric neurology*, vol. 28, no. 3, pp. 205–211, 2003.

[125] C. Loat, S. Curran, C. Lewis, J. Duvall, D. Geschwind, P. Bolton, and I. Craig, "Methyl-cpg-binding protein 2 polymorphisms and vulnerability to autism," *Genes, Brain and Behavior*, vol. 7, no. 7, pp. 754–760, 2008.

[126] Y.-h. Jiang, T. Sahoo, R. C. Michaelis, D. Bercovich, J. Bressler, C. D. Kashork, Q. Liu, L. G. Shaffer, R. J. Schroer, D. W. Stockton, *et al.*, "A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for ube3a," *American Journal of Medical Genetics Part A*, vol. 131, no. 1, pp. 1–10, 2004.

[127] A. Nguyen, T. A. Rauch, G. P. Pfeifer, and V. W. Hu, "Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, rora, whose protein product is reduced in autistic brain," *The FASEB Journal*, vol. 24, no. 8, pp. 3036–3051, 2010.

[128] C. Wong, E. L. Meaburn, A. Ronald, T. Price, A. Jeffries, L. Schalkwyk, R. Plomin, and J. Mill, "Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits," *Molecular psychiatry*, 2013.

[129] V. W. Hu, "From genes to environment: Using integrative genomics to build a "systems-level" understanding of autism spectrum disorders," *Child development*, vol. 84, no. 1, pp. 89–103, 2013.

[130] C. A. Pardo, D. L. Vargas, and A. W. Zimmerman, "Immunity, neuroglia and neuroinflammation in autism," *International review of psychiatry*, vol. 17, no. 6, pp. 485–495, 2005.

[131] S. Faber, G. M. Zinn, J. C. Kern Ii, and H. Skip Kingston, "The plasma zinc/serum copper ratio as a biomarker in children with autism spectrum disorders," *Biomarkers*, vol. 14, no. 3, pp. 171–180, 2009.

[132] S. Moore, P. Turnpenny, A. Quinn, S. Glover, D. Lloyd, T. Montgomery, and J. Dean, "A clinical study of 57 children with fetal anticonvulsant syndromes," *Journal of Medical Genetics*, vol. 37, no. 7, pp. 489–497, 2000.

[133] K. Strömland, V. Nordin, M. Miller, B. Akerström, and C. Gillberg, "Autism in thalidomide embryopathy: a population study," *Developmental Medicine & Child Neurology*, vol. 36, no. 4, pp. 351–356, 1994.

[134] H. Gardener, D. Spiegelman, and S. L. Buka, "Prenatal risk factors for autism: comprehensive meta-analysis," *The British journal of psychiatry*, vol. 195, no. 1, pp. 7–14, 2009.

[135] E. T. Parner, S. Baron-Cohen, M. B. Lauritsen, M. Jørgensen, L. A. Schieve, M. Yeargin-Allsopp, and C. Obel, "Parental age and autism spectrum disor-

ders," *Annals of epidemiology*, vol. 22, no. 3, pp. 143–150, 2012.

[136] A. M. Grabrucker, "Environmental factors in autism," *Frontiers in Psychiatry*, vol. 3, p. 118, 2013.

[137] E. Kolozsi, R. Mackenzie, F. Roullet, D. Decatanzaro, and J. Foster, "Prenatal exposure to valproic acid leads to reduced expression of synaptic adhesion molecule neuroligin 3 in mice," *Neuroscience*, vol. 163, no. 4, pp. 1201–1210, 2009.

[138] C. J. Phiel, F. Zhang, E. Y. Huang, M. G. Guenther, M. A. Lazar, and P. S. Klein, "Histone deacetylase is a direct target of valproic acid, a potent anticonvulsant, mood stabilizer, and teratogen," *Journal of Biological Chemistry*, vol. 276, no. 39, pp. 36734–36741, 2001.

[139] P. A. Main, M. T. Angley, P. Thomas, C. E. O'Doherty, and M. Fenech, "Folate and methionine metabolism in autism: a systematic review," *The American journal of clinical nutrition*, pp. ajcn–29002, 2010.

[140] C. M. Beard, L. A. Panser, and S. K. Katusic, "Is excess folic acid supplementation a risk factor for autism?," *Medical hypotheses*, vol. 77, no. 1, pp. 15–17, 2011.

[141] R. J. Schmidt, D. J. Tancredi, S. Ozonoff, R. L. Hansen, J. Hartiala, H. Allayee, L. C. Schmidt, F. Tassone, and I. Hertz-Picciotto, "Maternal periconceptional folic acid intake and risk of autism spectrum disorders and developmental delay in the charge (childhood autism risks from genetics and environment) case-control study," *The American journal of clinical nutrition*, vol. 96, no. 1, pp. 80–89, 2012.

[142] P. Surén, C. Roth, M. Bresnahan, M. Haugen, M. Hornig, D. Hirtz, K. K. Lie, W. I. Lipkin, P. Magnus, T. Reichborn-Kjennerud, *et al.*, "Association between maternal use of folic acid supplements and risk of autism spectrum disorders in children," *Jama*, vol. 309, no. 6, pp. 570–577, 2013.

[143] M. V. S. R. Group *et al.*, "Prevention of neural tube defects: results of the medical research council vitamin study," *The lancet*, vol. 338, no. 8760,

pp. 131–137, 1991.

[144] H. J. Blom, G. M. Shaw, M. den Heijer, and R. H. Finnell, "Neural tube defects and folate: case far from closed," *Nature Reviews Neuroscience*, vol. 7, no. 9, pp. 724–731, 2006.

[145] K.-c. Kim, S. Friso, and S.-W. Choi, "Dna methylation, an epigenetic mechanism connecting folate to healthy embryonic development and aging," *The Journal of nutritional biochemistry*, vol. 20, no. 12, pp. 917–926, 2009.

[146] C. L. Ulrey, L. Liu, L. G. Andrews, and T. O. Tollefsbol, "The impact of metabolism on dna methylation," *Human molecular genetics*, vol. 14, no. suppl 1, pp. R139–R147, 2005.

[147] M. Boris, A. Goldblatt, J. Galanko, S. J. James, *et al.*, "Association of mthfr gene variants with autism," *J Am Phys Surg*, vol. 9, no. 4, pp. 106–8, 2004.

[148] S. P. Paşca, E. Dronca, T. Kaucsár, E. C. Craciun, E. Endreffy, B. K. Ferencz, F. Iftene, I. Benga, R. Cornean, R. Banerjee, *et al.*, "One carbon metabolism disturbances and the c677t mthfr gene polymorphism in children with autism spectrum disorders," *Journal of cellular and molecular medicine*, vol. 13, no. 10, pp. 4229–4238, 2009.

[149] X. Zhao, A. Leotta, V. Kustanovich, C. Lajonchere, D. H. Geschwind, K. Law, P. Law, S. Qiu, C. Lord, J. Sebat, *et al.*, "A unified genetic theory for sporadic and inherited autism," *Proceedings of the National Academy of Sciences*, vol. 104, no. 31, pp. 12831–12836, 2007.

[150] M. Ronemus, I. Iossifov, D. Levy, and M. Wigler, "The role of de novo mutations in the genetics of autism spectrum disorders," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 133–141, 2014.

[151] A. Purcell, O. Jeon, A. Zimmerman, M. Blue, and J. Pevsner, "Postmortem brain abnormalities of the glutamate neurotransmitter system in autism," *Neurology*, vol. 57, no. 9, pp. 1618–1628, 2001.

[152] V. W. Hu, B. C. Frank, S. Heine, N. H. Lee, and J. Quackenbush, "Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes," *BMC genomics*, vol. 7, no. 1, p. 118, 2006.

[153] Y. Nishimura, C. L. Martin, A. Vazquez-Lopez, S. J. Spence, A. I. Alvarez-Retuerto, M. Sigman, C. Steindler, S. Pellegrini, N. C. Schanen, S. T. Warren, *et al.*, "Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways†," *Human molecular genetics*, vol. 16, no. 14, pp. 1682–1698, 2007.

[154] K. Garbett, P. J. Ebert, A. Mitchell, C. Lintas, B. Manzi, K. Mirnics, and A. M. Persico, "Immune transcriptome alterations in the temporal cortex of subjects with autism," *Neurobiology of disease*, vol. 30, no. 3, pp. 303–311, 2008.

[155] J. P. Gregg, L. Lit, C. A. Baron, I. Hertz-Picciotto, W. Walker, R. A. Davis, L. A. Croen, S. Ozonoff, R. Hansen, I. N. Pessah, *et al.*, "Gene expression changes in children with autism," *Genomics*, vol. 91, no. 1, pp. 22–29, 2008.

[156] A. M. Enstrom, L. Lit, C. E. Onore, J. P. Gregg, R. L. Hansen, I. N. Pessah, I. Hertz-Picciotto, J. A. Van de Water, F. R. Sharp, and P. Ashwood, "Altered gene expression and function of peripheral blood natural killer cells in children with autism," *Brain, behavior, and immunity*, vol. 23, no. 1, pp. 124–133, 2009.

[157] V. W. Hu, A. Nguyen, K. S. Kim, M. E. Steinberg, T. Sarachana, M. A. Scully, S. J. Soldin, T. Luu, and N. H. Lee, "Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis," *PloS one*, vol. 4, no. 6, p. e5775, 2009.

[158] M. D. Alter, R. Kharkar, K. E. Ramsey, D. W. Craig, R. D. Melmed, T. A. Grebe, R. C. Bay, S. Ober-Reynolds, J. Kirwan, J. J. Jones, *et al.*, "Autism and increased paternal age related changes in global levels of gene expression regulation," *PloS one*, vol. 6, no. 2, p. e16715, 2011.

[159] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[160] S. Kong, Y. Shimizu-Motohashi, M. Campbell, I. Lee, C. Collins, S. Brewster, I. Holm, L. Rappaport, I. Kohane, and L. Kunkel, "Peripheral blood gene expression signature differentiates children with autism from unaffected siblings," *neurogenetics*, vol. 14, no. 2, pp. 143–152, 2013.

[161] J. T. Bell and T. D. Spector, "A twin approach to unraveling epigenetics," *Trends in Genetics*, vol. 27, no. 3, pp. 116–125, 2011.

[162] J. T. Bell and R. Saffery, "The value of twins in epigenetic epidemiology," *International Journal of Epidemiology*, p. dyr179, 2012.

[163] M. Ketelaar, R. Hofstra, and M. Hayden, "What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration," *Clinical genetics*, vol. 81, no. 4, pp. 325–333, 2012.

[164] J. Van Dongen, P. E. Slagboom, H. H. Draisma, N. G. Martin, and D. I. Boomsma, "The continuing value of twin studies in the omics era," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 640–653, 2012.

[165] K. Kim, K. Lee, H. Bang, J. Y. Kim, and J. K. Choi, "Intersection of genetics and epigenetics in monozygotic twin genomes," *Methods*, 2015.

[166] I. I. Gottesman and A. Bertelsen, "Confirming unexpressed genotypes for schizophrenia: Risks in the offspring of fischer's danish identical and fraternal discordant twins," *Archives of General Psychiatry*, vol. 46, no. 10, pp. 867–872, 1989.

[167] C. Kakiuchi, K. Iwamoto, M. Ishiwata, M. Bundo, T. Kasahara, I. Kusumi, T. Tsujita, Y. Okazaki, S. Nanko, H. Kunugi, *et al.*, "Impaired feedback regulation of xbp1 as a genetic risk factor for bipolar disorder," *Nature genetics*, vol. 35, no. 2, pp. 171–175, 2003.

[168] R. Plomin and D. Daniels, "Why are children in the same family so different from one another," *Behavioral and Brain Sciences*, vol. 10, no. 1, pp. 1–16, 1987.

[169] A. Pike, D. Reiss, E. M. Hetherington, and R. Plomin, "Using mz differences in the search for nonshared environmental effects," *Journal of Child Psychology and Psychiatry*, vol. 37, no. 6, pp. 695–704, 1996.

[170] J. T. Dudley, R. Tibshirani, T. Deshpande, and A. J. Butte, "Disease signatures are robust across tissues and experiments," *Molecular systems biology*, vol. 5, no. 1, p. 307, 2009.

[171] V. G. Cheung, L. K. Conlin, T. M. Weber, M. Arcaro, K.-Y. Jen, M. Morley, and R. S. Spielman, "Natural variation in human gene expression assessed in lymphoblastoid cells," *Nature genetics*, vol. 33, no. 3, pp. 422–425, 2003.

[172] Y. H. Yang and T. P. Speed, "Design and analysis of comparative microarray experiments," *Statistical analysis of gene expression microarray data*, pp. 35–91, 2003.

[173] S. Pounds and C. Cheng, "Sample size determination for the false discovery rate," *Bioinformatics*, vol. 21, no. 23, pp. 4263–4271, 2005.

[174] P. Liu and J. G. Hwang, "Quick calculation for sample size while controlling false discovery rate with application to microarray analysis," *Bioinformatics*, vol. 23, no. 6, pp. 739–746, 2007.

[175] P. M. Visscher and D. Posthuma, "Statistical power to detect genetic loci affecting environmental sensitivity," *Behavior genetics*, vol. 40, no. 5, pp. 728–733, 2010.

[176] Z. Kaminsky, A. Petronis, S.-C. Wang, B. Levine, O. Ghaffar, D. Floden, and A. Feinstein, "Epigenetics of personality traits: an illustrative study of identical twins discordant for risk-taking behavior," *Twin Research and Human Genetics*, vol. 11, no. 01, pp. 1–11, 2008.

[177] P.-C. Tsai and J. T. Bell, "Power and sample size estimation for epigenome-wide association scans to detect differential dna methylation," *International journal of epidemiology*, vol. 44, no. 4, pp. 1429–1441, 2015.

[178] C.-C. Liew, J. Ma, H.-C. Tang, R. Zheng, and A. A. Dempsey, "The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool," *Journal of Laboratory and Clinical Medicine*, vol. 147, no. 3, pp. 126–132, 2006.

[179] C. Cai, P. Langfelder, T. F. Fuller, M. C. Oldham, R. Luo, L. H. van den Berg, R. A. Ophoff, and S. Horvath, "Is human blood a good surrogate for brain tissue in transcriptional studies?," *BMC genomics*, vol. 11, no. 1, p. 1, 2010.

[180] M. N. Davies, S. Lawn, S. Whatley, C. Fernandes, R. W. Williams, and L. C. Schalkwyk, "To what extent is blood a reasonable surrogate for brain in gene expression studies: estimation from mouse hippocampus and spleen," *Frontiers in neuroscience*, vol. 3, p. 2, 2009.

2 Gene expression profiling of monozygotic twins with autism spectrum disorder

2.1 Introduction

Autism spectrum disorder (ASD) is perhaps unique amongst neuropsychiatric conditions in that in addition to the etiology being largely unaccounted for, there is also no unified pathomechanistic explanation. Complicating matters somewhat is the proposed role of non-genetic factors, as evidenced by incomplete concordance in monozygotic (MZ) twins, which could be contributing to liability as well as obfuscating any underlying genetic susceptibility. Offering a complementary route forward to genetic studies, are functional genomics approaches which profile genome-wide patterns of gene expression and methylation. Such studies must contend with the confounding influence of genetics, along with the possibility that molecular differences could have arisen as a secondary consequence of the disorder [1]. To help circumvent some of these issues, it has been suggested utlizing a discordant MZ twin-based study design [2, 3, 4, 5, 1]. Since MZ twins are matched for age, sex, and genetic background, as well as some early environmental influences [5], in theory any divergence in molecular profiles can be considered as the driver of the observed phenotypic discordance [1]. It is hoped therefore that assessing the molecular basis of twin discordance could help to separate genetic from non-genetic factors, identify molecular signatures in the form of altered gene expression

or methylation profiles, and ultimately elucidate disorder-related pathways.

Genome-wide gene expression profiling of MZ twins has previously been performed for a range of psychiatric and neurological conditions including schizophrenia [6], bipolar disorder [7], and Huntington's disease [8]. There has however, been limited application of this methodology to ASD; we were only able to identify one such study by Hu et al., which examined gene expression in 3 pairs of MZ ASD discordant and two pairs of MZ ASD concordant twins, identifying differences in gene networks involved in neurological function, nervous system development and inflammation [9]. Interestingly, several studies have utilised MZ twins to investigate DNA methylation patterns in ASD (see Chapter 1). In what was at the time the largest study of its kind, our group previously characterised genome-wide methylation patterns in a cohort of 50 MZ twin pairs (100 individuals) that included ASD concordant, ASD discordant pairs, as well as unaffected concordant controls. The analysis strategy involved measuring differences within discordant pairs and between groups (cases and controls) [10]. The main findings from the within-discordant pairs comparison were that while ASD was not associated with large-scale disruptions in methylation, a number of CpG sites were found to vary in methylation levels between co-twins, with the top result a site in the *NFYC* gene - found to be consistently hypermethylated in affected compared to unaffected co-twins across all pairs. Between-groups analysis of cases and controls identified additional ASD-associated differentially methylated sites. Overall, ASD-associated differential methylation was observed in the vicinity of a number of genes previously implicated in the disorder including: *AFF2, AUTS2, GABRB3, NLGN3, NRXN1, SLC6A4*.

The MZ differences approach has thus far been under utilised in molecular studies of ASD, and we are also unaware of any single study that has explored both gene expression and DNA methylation differences in the same sample. Here, we address this perceived gap in the literature by performing a follow-up to the previous DNA methylation study by Wong et al., using the same cohort of ASD concordant, ASD discordant, ASD trait discordant, and unaffected concordant control MZ twin pairs. Genome-wide gene expression profiles for each subject are first generated from whole blood samples using microarray technology, and two main

analyses carried out. Firstly, a between-groups analysis of affected and unaffected individuals, to identity disrupted genes and pathways that are common between both concordant and discordant pairs. Secondly, we perform a within-pairs analysis of discordant pairs alongside a within-group discordant analysis, in order to identify perturbations arising from non-genetic factors and identify those in common between discordant twin pairs. The overall aims of the study are twofold, firstly, to uncover patterns of gene expression associated with ASD in both concordant and discordant pairs, in order to provide convergent evidence for genes and pathways disregulated in ASD. And secondly, by investigating expression differences common in MZ discordant pairs, identify those genes and networks which are potentially sensitive to the environment. It is hoped that the results will provide further support for the findings from the previous methylation study, as well as providing a high quality dataset for a planned integrative study aiming to establish the functional relevance of identified signals.

2.2 Methods

2.2.1 Subjects

The subjects used in this study are a subsample from the Twins Early Development Study (TEDS), a longitudinal study investigating the cognitive and behavioural development of twins born in the UK between January 1994 and December 1996 [11, 12, 13]. Ethical approval for the original study was provided by the Institute of Psychiatry, Psychology and Neuroscience (IoPPN) ethics committee, reference number 05/Q0706/228.

Participating individuals completed various web and telephone-based tests and questionnaires at regular intervals over childhood and adolescence designed to assess various aspects of cognition, language and behaviour (see [13] for further details). As part of this, twins were assessed for ASD related traits and behaviours at ages 8 and 12 using the Childhood Aspergers Symptom Test (CAST). The CAST is a 31 item screening test completed by parents and teachers in a non-clinical setting,

which assesses ASD based on the DSM-IV [14] defined subscales of social symptoms (12 items), non-social symptoms (7 items), and communication impairments (12 items)[15, 16]. The subscale measurements are combined additively to give a total CAST score out of 31, with those scoring ≥ 15 categorised as "at risk". Individuals identified as at risk were also formally assessed at home using the Autism Diagnostic Interview-Revised (ADIR) [17] and the Autism Diagnostic Observation Schedule (ADOS) [18], both considered gold standard diagnostic tools.

2.2.2 Subject selection and study groups

Study group	Classification	ADI-R/ADOS	CAST	Gender (M:F)	Sample size
1	Concordant ASD	Both	Total CAST ≥ 15	12:0	12
2	Discordant ASD	Affected	N/A	6:6	12
3	Discordant social CAST	Neither	δ social CAST ≥ 2 S.D.	6:14	20
4	Discordant communication CAST	Neither	δ communication CAST ≥ 2 S.D	12:8	20
5	Discordant non-social CAST	Neither	δ non-social CAST ≥ 2 S.D.	14:6	20
6	Concordant low CAST	Neither	Total CAST ≤ $\bar{x}$	8:14	22
					106

Table 2.1. Subjects were assigned to one of six study groups based on ADI-R/ADOS diagnosis and/or CAST scores. The ADI-R/ADOS column refers to which co-twin(s) received the diagnosis, either both, the affected twin only, or neither. In the CAST column, δ refers to the difference in scores between co-twins on each of the scales, where S.D. is the standard deviation and $\bar{x}$ is the sample mean.

A total of 53 MZ twin pairs (106 individuals) were selected from TEDS. Selections were based on scores on the CAST subscales and total score, and whether a clinical diagnosis of ASD had been made. Study groups were defined and the twins assigned as follows: group 1 : 6 ASD concordant pairs (both members of the pair having had a formal ASD diagnosis), group 2 : 6 ASD discordant pairs (one member of the pair with a formal ASD diagnosis), groups 3,4,5 : each consisting of 10 ASD trait discordant pairs for each of the CAST subscales: social, non-social, communication (co-twins score greater than two standard deviations apart), and group 6 : 11 unaffected concordant control pairs (both members of twin pair scored less than or equal to the sample mean in total CAST score). A summary of the study groups is provided in Table 2.1 and represented visually in Figure 2.1.

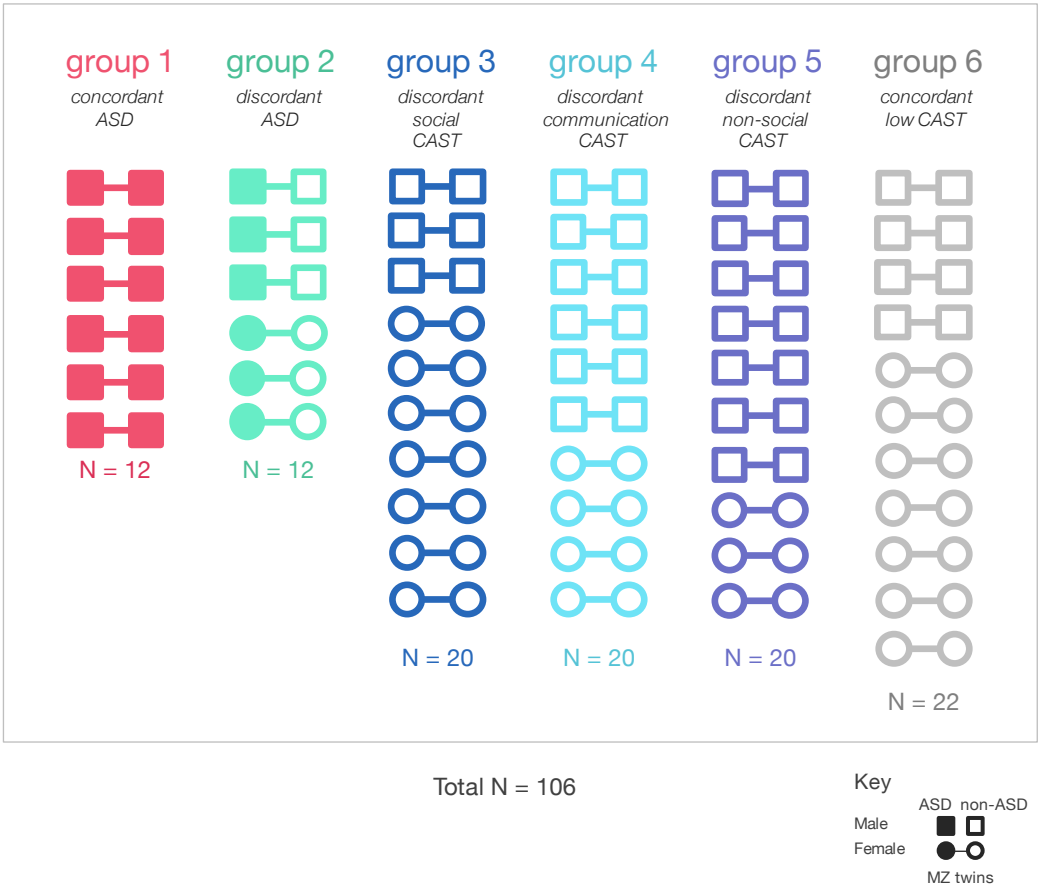

Figure 2.1. The study sample - with group, sex, pair and ASD affected status indicated.

2.2.3 Sample collection

Participating families visited the IoPPN when the twins were aged 15 to have whole blood samples taken by a trained phlebotomist. Using the Ambion PAXgene RNA system (PreAnalytiX, QIAGEN, Germany), three 2.5 ml tubes of blood were collected and stored. In addition, 3 ml of blood was collected in an EDTA tube to assess whole blood cell subtype composition, which for all samples was found to be within normal ranges.

2.2.4 HuGe microarray

Microarray technology was used to produce genome-wide measurements of gene expression for the subjects. These experiments were performed at the IoPPN, using the Affymetrix Human Gene 1.0 ST (HuGe) platform. The HuGe array is designed to measure the expression of 32,020 well-annotated RefSeq coding transcripts mapping to a total of 21,014 Entrez genes, and does so using 25-mer probes distributed across transcribed regions (median 26 probes per gene). Total RNA was isolated and extracted from the PAXgene blood tube using the PAXgene blood RNA kit (PreAnalytiX, QIAGEN, Germany), and globin related transcripts removed using the Ambion GLOBINclear (Ambion,USA) kit. These samples were then run on HuGe arrays (one per sample) following the manufacturer's instructions: sense strand cDNA was generated using the Ambion WT Expression Kit (Affymetrix, USA), fragmented and labeled using the AffyMetrix GeneChip WT Terminal Labelling Kit (Affymetrix, USA), and finally the prepared samples were hybridized to the arrays and scanned using the GeneChip Scanner 3000 7G (Affymetrix, USA).

All of data processing and analysis steps described below were performed in the *R* statistical environment [19], following the analysis workflow described in the *oligo*[20] package in *bioconductor* [21].

2.2.5 Quality control I - microarray data

2.2.5.1 Inspection of array images

The data generated by the arrays was subject to a series of quality control steps. To begin with, following the manufacturer's guidelines, the scanner images (.DAT files) were manually inspected to check for spatial intensity artifacts, such as overly bright spots, or other irregularities like scratches or air bubbles. In this instance, no issues were detected.

2.2.5.2 Distribution of raw intensity measurements

Plots were made of the average raw intensity and mean absolute deviation of residuals for each array to help identify outlying arrays (Figure 2.2 A and B). Examining these plots, it was observed that arrays 34 (affected male from group 4) and 37 (affected female from group 5) were dissimilar to the others, displaying lower average signal intensity and greater variance in intensity measurements. For these reasons, these arrays were removed from the dataset, leaving expression data for 104 individuals (51 complete twin pairs) to take forward for pre-processing.

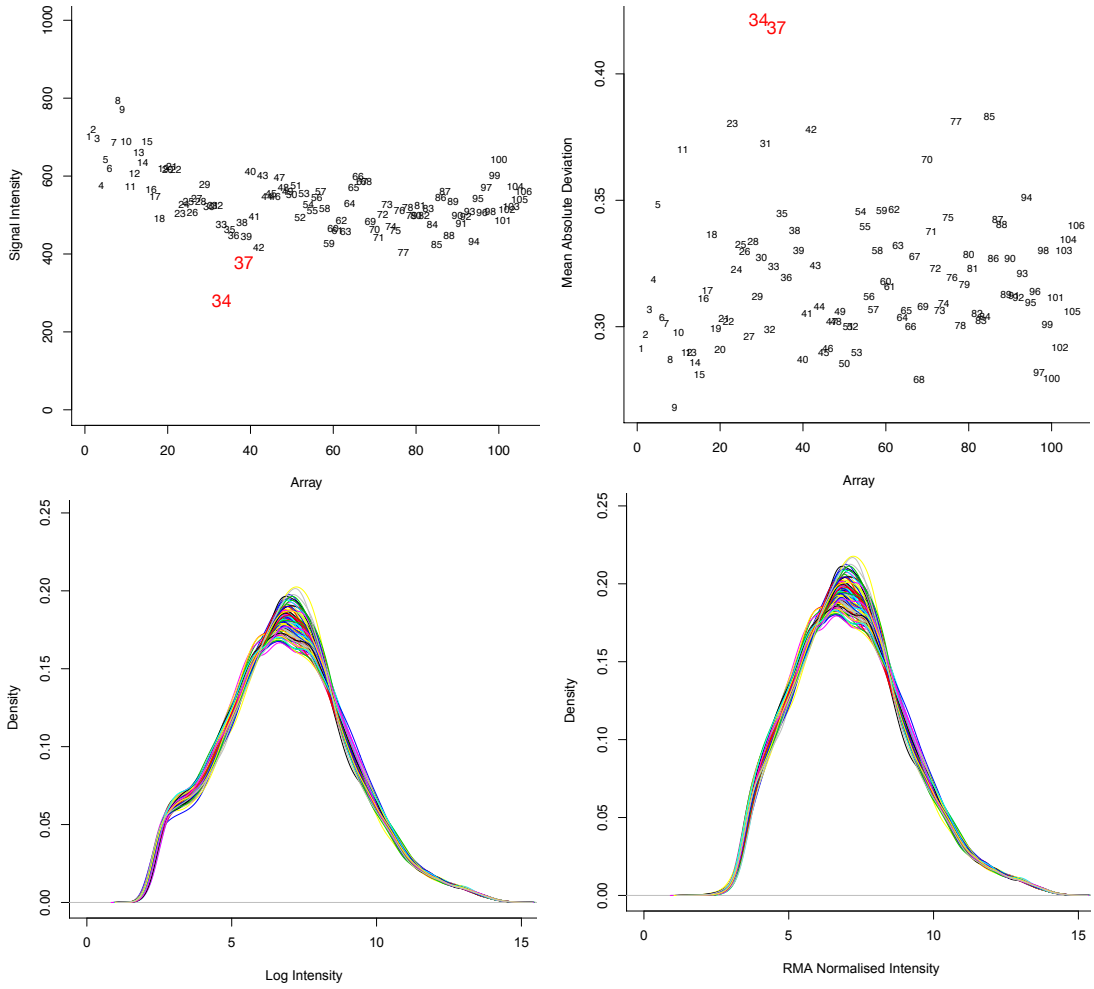

Figure 2.2. Plots showing a. the average raw intensity, b. mean absolute deviation of residuals for each array. Based on these plots, arrays 34 and 37 were both identified as outliers and removed from the dataset. c. distribution of log intensity signals for each array prior to normalization, d. distribution of intensities after RMA normalization. These plots show the effect of normalisation in bringing the intensity distributions for each array onto a common scale.

2.2.6 Data pre-processing

Probeset intensity measurements (.CEL files) were processed using the oligo package, alongside the pd.hugene.1.0.st.v1 [22] annotation package. Next, RMA (robust multi-array analysis) [23], as implemented in *oligo*, was used to summarise the probe level data (for core probesets targeting RefSeq annotated transcripts), apply background correction, and perform quantile to quantile normalisation, before finally log-transforming the intensity values. The resulting expression dataset contained normalised, log-transformed expression values for 33,297 probesets.

2.2.7 Quality control II - expression data set

2.2.7.1 Probeset filtering / removal of background signal

To gain an impression of data quality, the overall distribution of probeset intensity measurements across all of the arrays was inspected. This revealed a non-normal distribution with a large and small peak (Figure 2.2 C). The smaller peak was believed to represent low-signal probesets and potential background noise on the arrays; a well known technical issue in array expression studies. For Affymetrix arrays based on the Human Exon 1.0 ST design, the DABG (detection above background) method (as implemented in Affymetrix power tools command-line software) is recommended for filtering out low-signal probesets. This works by comparing probesets to a set of control probes, yielding detection p-values. For the Human Gene 1.0 ST array however, which uses only a subset of the probes on the Exon array to perform gene-level profiling, DABG is not considered to be a robust metric [24]. While other simple filtering strategies can be used to remove genes not detected above background, such as removing probes that do not meet the overall mean intensity in at least 50% of the samples, a general consensus is lacking. Based on the previous observations about the overall distribution of intensities, a different approach was taken, which sought to model the intensities as a mixture of two different distributions: one representing the low-signal/background and one the true signal, and then filtering out probes with intensity profiles matching the back-

ground. To this end, the *normalmixEM* function from the *mixtools* package [25] in *R* was used. Briefly, the *normalmixEM* function uses the expectation maximization (EM) algorithm to iteratively fit parameters to a finite mixture model with the goal of maximizing the expectation for a predefined number of functions, a form of maximum likelihood estimation (MLE) - see [25] for more details. For each data point, a likelihood is estimated that it has been generated by each of the functions in the mixture. This procedure was run on the processed expression data, and any probeset with greater than 0.8 likelihood of belonging to the smaller distribution was filtered out, resulting in the removal of 2273 low-signal probesets (Figure 2.2 D). The code is provided in Appendix - Chapter 2 - Figure A1.

2.2.7.2 Probeset aggregation

Next, probesets were mapped to gene annotation using the biomaRt package [26], and those mapping to the same gene were aggregated, retaining only the intensity readings of the probeset with the greatest variance across all of the samples. This left a total of 26,482 probesets for core analysis.

2.2.8 Exploratory analysis

The quality of the final expression dataset was assessed. Given the high heritability of gene expression, co-twins (as genetically related individuals) would be expected to show greater similarity in measures of transcript abundance than unrelated individuals. Similarity was assessed by two different means. Firstly, the correlation between expression profiles for all subjects was calculated using Pearson's *r*, as implemented in the *cor* function in *R*:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}) \tag{2.1}$$

Secondly, the Euclidean distance for all pairwise comparisons was calculated :

$$d = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{2.2}$$

across all m dimensions (probes) for samples x and y. The data was then clustered (hierarchical, unsupervised, complete linkage method), using the *dist* and *hclust* functions in *R*. Clustering was performed firstly with the entire set of probes, and then with a subset of the top 100 sex specific transcripts published by the WEHI bioinformatics group [27], which includes genes on the X chromosome reported to escape X inactivation [28] and a set of Y chromosome genes [29]), to examine clustering based on sex. The numerical data and resulting heatmaps /dendrograms were manually inspected.

2.2.9 Differential expression analysis

2.2.9.1 T-test and Wilcoxon

Before performing the main differential expression analysis, to gain an initial impression of the extent and magnitude of the differences in expression levels being detected, traditional parametric and non-parametric inferential testing methods were used. For the comparison of affected (ASD cases) and unaffected individuals across groups 1,2, and 6 (concordant ASD, discordant ASD, concordant controls) the independent two sample Student's t-test was used:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.3}$$

where $\bar{X}$ is the mean, n the sample size, and s_p the pooled standard deviation for the two experimental conditions (i.e. ASD affected vs unaffected). The Wilcoxon rank sum (or Mann-Whitney U) test was also used as the non-parametric equivalent:

$$min(U_1 = R_1 - \frac{n_1(n_1+1)}{2}, U_2 = R_2 - \frac{n_2(n_2+1)}{2}) \tag{2.4}$$

taking the minimum of the two U statistics, where R is the sum of ranks after all observations have been ranked from smallest to largest, and n the sample size for each of the two conditions.

For the between pairs comparison of affected and unaffected individuals from group 2 (discordant ASD) was also performed using the paired version of the two sample t-test:

$$t = \frac{\bar{X}_D - \mu_0}{\frac{s_D}{\sqrt{n}}} \tag{2.5}$$

where X_D is the mean differences between case and control within the pairs, $\mu_0 = 0$ the mean of the population differences, s_D the standard deviation of the differences, and n the number of pairs.

And for the non-parametric equivalent, the Wilcoxon matched pairs signed ranks test was used:

$$W = \sum_{i=1}^{n_r}(sgn(x_{2,i} - x_{1,i}) \times R_i) \tag{2.6}$$

which is the sum of the signed ranks for n_r the sample size of pairs with non-zero differences, where R_i is the rank of pair (ordered by the magnitude of the difference).

2.2.9.2 Regression analysis

It is widely acknowledged that statistics such as the t-statistic do not perform well in genome-wide expression profiling experiments, and can lead to an increase in the number of false positives [30, 31, 32]. One reason for this is that estimates of variance (i.e. the denominator in the t-test) can be unstable, due to each gene on the microarray having a different variance [32].

Primarily for these reasons, for the main analysis the *limma* method was used [33] (part of the *bioconductor* for *R* suite [21]).This method works by fitting a linear model to the expression data for each gene, and testing for association with the trait using a moderated t statistic. This approach offers several advantages over classic tests. Firstly, it uses an empirical Bayes approach to borrow information across genes to derive stable gene specific variance estimates [34]; particularly advantageous for experiments with small numbers of replicates, as information borrowing can increase power to detect differentially expressed genes [33]. Secondly, a regression framework allows for the inclusion of multiple covariates into the model, which can help to separate out the systematic variation from the biological. And thirdly, related to the previous point, experimental designs using biological and technical replicates are supported, relevant here because of the use of paired data.

Two major analyses were planned. To begin with, a between-groups comparison (case-control) would be made using ASD affected and unaffected individuals across groups 1,2, and 6 (concordant ASD, discordant ASD, concordant controls). This was intended to capture trait-associated expression differences common to concordant and discordant cases when compared to age-matched controls, attributable to genetic and environmental factors. Next, within-group analyses would be carried out for the discordant groups 2,3,4,5 (discordant ASD, discordant social CAST, discordant communication CAST, discordant non-social CAST), to characterise ASD-associated differences attributable to non-shared environmental factors.

Linear model selection: A linear model describes the relationship between an outcome of interest (the dependent variable) and one or more predictor variables. The expression values y for j genes in i subjects as explained by n variables can be modelled as follows:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1i}...\beta_{nj}x_{ni} + \epsilon_{ij} \tag{2.7}$$

where β_0 the intercept term, plus effects for n explanatory variables (β_1 to β_n), and finally an error term ϵ. In limma, this is represented in vector form as:

$$Y_j = X\beta_j + \epsilon_j \tag{2.8}$$

where Y is a vector of the expression values for gene j, X is a design matrix with rows representing samples and columns representing the different experimental variables/treatments, β a vector of coefficients, and ϵ the vector of error terms.

Linear regression uses the ordinary least squares method to estimate values for the beta coefficients, with the objective to minimize the sum of the squared residuals for the observations. In limma, each gene is regressed on the product of the design matrix by the coefficients β.

A number of different models were initially considered. In a hypothetical, fully balanced version of the experiment, with every study group containing both males and females and complete twin pairs, the following single combined, mixed model could conceivably be used:

$$expression \sim 0 + case + group + (1|pair) + (case : group) \tag{2.9}$$

where 0 indicates that there is no intercept term, group is a reduced dimension version of study group status where groups are classed as concordant or discordant, case is the affected status, 1 | pair is the random effect of pair (effects that are not constant across the sample, in limma, achieved through the use of the duplicate correlation method [35] to model intrapair correlation). The omission of the intercept term allows multiple different contrasts to be fit, as there is no longer a baseline level of comparison. An experimental contrast of interest α_j can then be obtained using a contrast matrix C and coefficients β_j from above:

$$\alpha_j = \beta_j C^T \tag{2.10}$$

Such an approach would be advantageous as it uses the entire dataset for estimating co-efficients, whilst allowing a variety of between-group and within-group comparisons to be made. For example, a case vs control analysis across all subjects could be performed by contrasting concordant:case and discordant:case with concordant:control and discordant:control, and a within-discordant group comparison could be made by contrasting discordant:case with discordant:control. The within-group analyses of the remaining discordant groups, as they do not contain any cases (only sub-clinical individuals with a high score in one of the CAST subscales), would have to be performed using a separate model, but through a similar formulation.

Unfortunately, in our dataset, because group, sex and case status are either nested or partially correlated measures (an inherent property of the sample), and also due to the presence of incomplete pairs, this approach could ultimately not be taken. Instead, individual models for case-control and within-group were used.

Case-control analysis: To assess ASD associated gene-expression difference common in both concordant and discordant cases, a between pairs case-control analysis was performed by fitting the following linear model:

$$expression \sim pair + case \tag{2.11}$$

whereby gene expression is regressed on twin pair and case status. Here, twin pair is a categorical, fixed effect and case a binary fixed effect. Group is omitted from the model due to potential confounding with case status, and sex has also not been included due to nesting within the pair effect (since MZ co-twins are the same sex).

Within-group analyses: Next, for discordant groups 2,3,4,5 (discordant ASD, discordant social CAST, discordant communication CAST, discordant non-social CAST) the above model was once again used to perform within-group analyses.

Male-only case-control and male-only within-group: Both case-control and within-group analyses were also performed with only male subjects to explicitly control for sex. This was carried out because of the potential for confounding due to sex-based differences in expression and uncertainty in the effectiveness of the linear models in removing these effects. Given the observed sex-bias in ASD [36], there is also a biological rationale for considering male and female cases separately, as this represents another possible source of heterogeneity.

sex analysis: Additionally, for quality control purposes and also to serve as a benchmark for comparative purposes, a sex-based, male-female contrast was made:

$$expression \sim sex \tag{2.12}$$

Assessing fit: The distribution of p-values from inferential tests for differential expression would be expected to be approximately uniform, as the majority of probes will not be associated with the trait. Should this assumption not be met, then attempting to establish a threshold for significance becomes essentially meaningless [37, 38]. Examining p-value distributions can therefore be helpful for detecting experimental and analytical issues. Extreme deviation from the null can be indicative of technical issues, such as probe cross-hybridization (non-specific binding to similar non-target sequences[39]) or sample contamination. The distribution of p-values can also be informative about the effectiveness of the overall analysis strategy used: imperfect normalization and poor variance estimation have been shown to lead to inaccurate p-value estimation [37]. To this end, quantile-quantile (QQ) plots of observed p distribution against expected p distributions were created and manually inspected for each of the tests.

Statistics for differential expression: The output generated by *limma* is a list of genes with associated test statistics including $\log_2$ fold change (logFC), moderated t statistic, p-value (calculated from the moderated t), and the false discovery rate (FDR) statistic (multiple testing adjusted p-value - see Chapter 4 section 1.5). In

this case, FDR was used to rank genes by the level of evidence for differential expression. A cut-off of 0.2 was defined as the threshold for significance, meaning that on average,for every four true positive findings, one false positive would be expected. This more permissive threshold was selected due to the exploratory nature of the study and the small sample size.

2.3 Results

2.3.1 Exploratory analysis

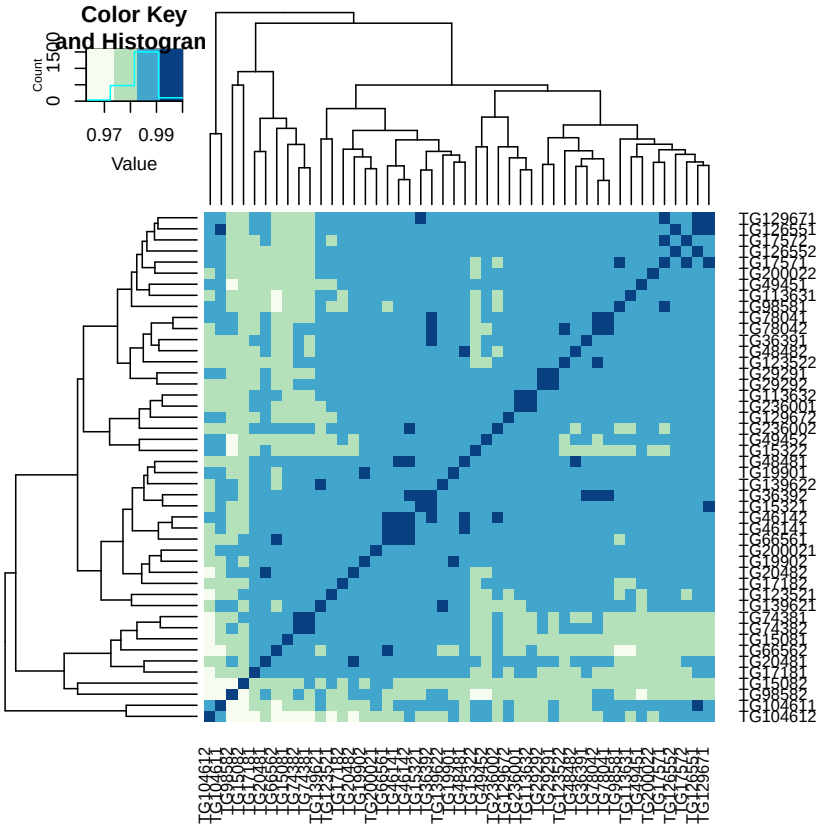

Figure 2.3. Heatmap showing correlations between individuals. High levels of correlation, r values of 0.99 and above, are indicated by the dark blue coloured squares. With high quality expression data, the expectation would be that co-twins within pairs would be more highly correlated, and should mainly cluster together. Furthermore, similarity to unrelated individuals would be expected to be lower, and the highest levels of correlation found within twin pairs, and so dark blue (high correlation) would be expected to run along the diagonal. However, here it can be seen that there are very few twin pairs that actually conform to this pattern, and furthermore, the sporadic patches of dark blue off the diagonal also indicate individuals who are highly correlated with several unrelated individuals.

Exploring the measures of sample similarity, the median correlation r between co-twins was 0.99 and for unrelated individuals slightly lower at 0.98. This is repre-

sented visually in the heatmap (Figure 2.3). It was observed that pairs of co-twins clustered together in nodes for only approximately half (56%) of the sample. The dendrogram is shown in Figure 2.4. Clustering based on sex-specific genes showed male and female subjects segregating into two distinct clusters, this is shown in Figure 2.5).

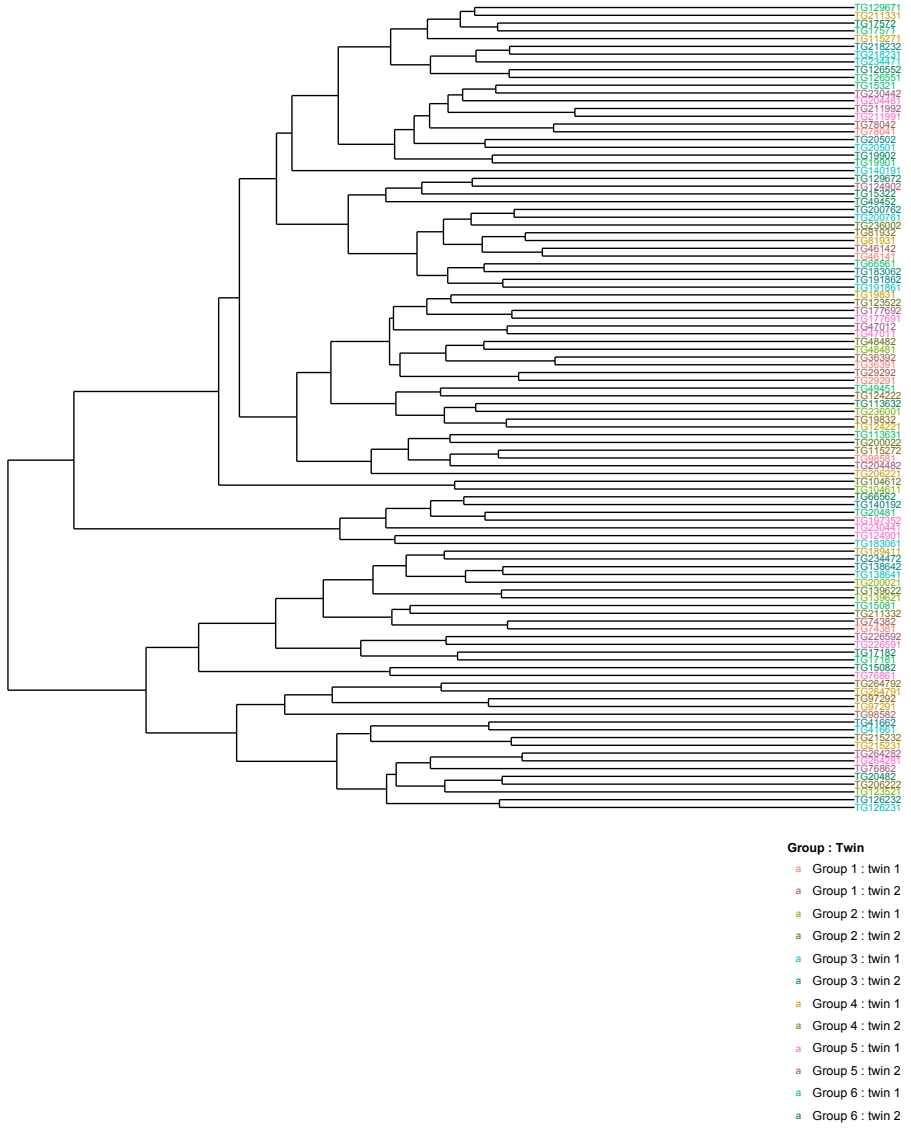

Figure 2.4. Dendrogram showing similarity of samples based on genome-wide expression values. Labels are coloured by group membership and co-twin number. Twins clustering together in pairs are concordant for hue and discordant for shade, and cases of unrelated individuals clustering together are discordant for hue. In this instance only around 56% of co-twins are found to cluster together as expected.

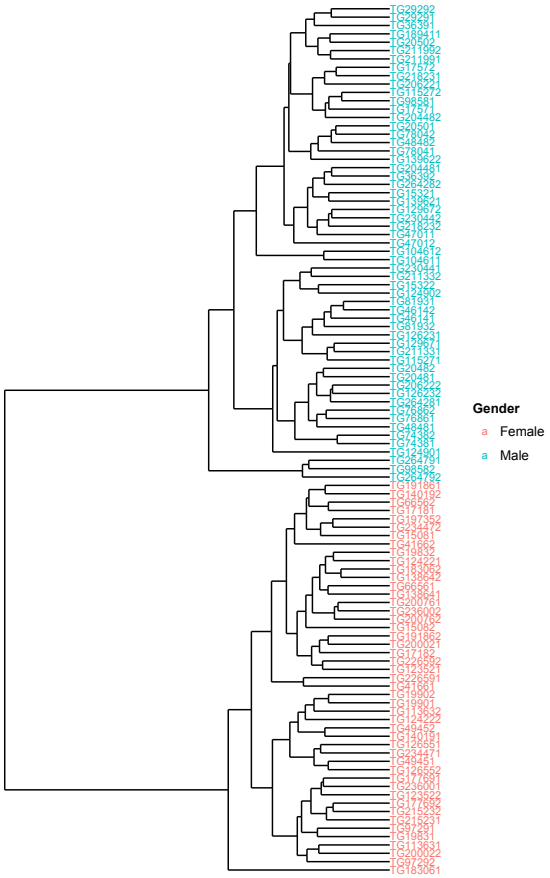

Figure 2.5. Dendrogram showing similarity of samples based on expression values for genes known to differ in expression levels between sexes. Labels are coloured by sex. Here it can be seen that males and females form two distinct clusters, indicating correct labelling by sex for the samples.

2.3.2 Differential expression analysis

2.3.2.1 Student's t and Wilcoxon

Student's t and Wilcoxon tests were first run and the QQ plots showing the resulting p-value distributions inspected (not included here). Neither method produced the expected null distribution, instead showing uniform deflation across the range of

p values. The same was also observed for each of the linear models - see Figure 2.6. Once again, these were unexpected findings with important implications for the interpretation of the results produced (see Discussion).

2.3.2.2 Regression

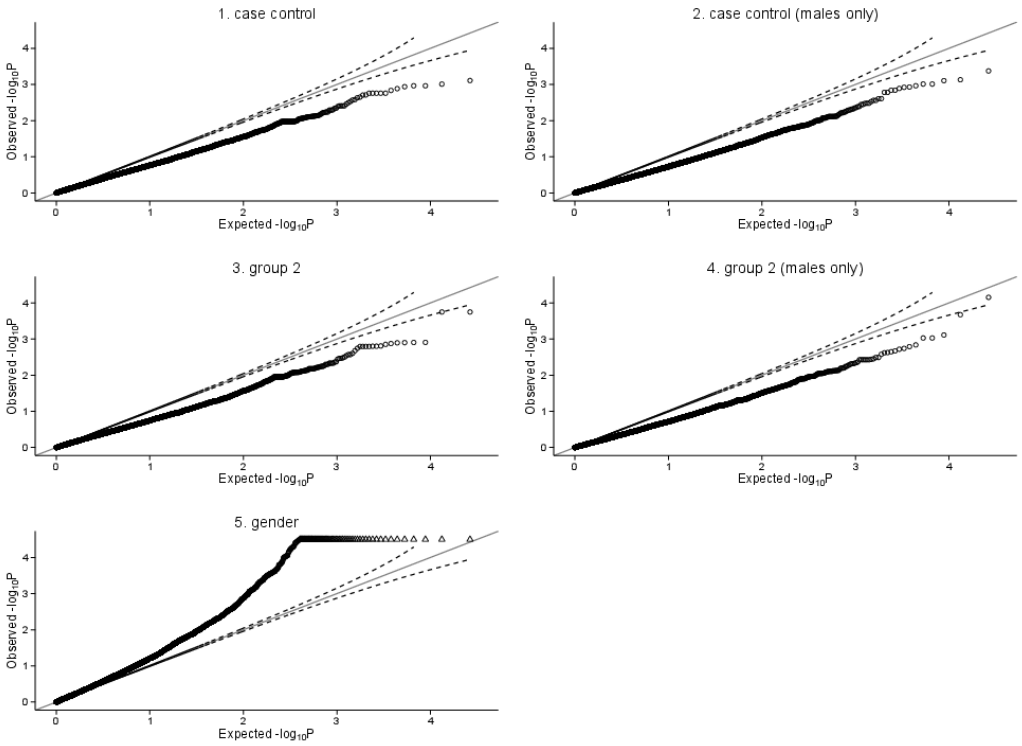

Figure 2.6. QQ plots showing observed vs theoretical distribution of p-values for each of the fitted models

The top 50 differentially expressed genes from the case-control analysis of groups 1,2, and 6 are shown in Table 2.2. Overall, the fold changes appear small and the associated p-values are not indicative of any highly significant findings. The results also contain a number of small nuclear RNAs (snRNA or U-RNA), making up an unexpectedly large proportion of the top 50 results.

The top 50 differentially expressed genes from the within-group 2 analysis of discordant pairs are presented in Table 2.3. As with the case-control results, the log

fold changes are small and p-values not indicative of highly significant findings. These issues are covered in more detail in the Discussion section.

Case-control

ENS_gene	hgnc_symbol	chr	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000154188	ANGPT1	8	0.29	5.8	3.8	0.00078	1	-2.1
ENSG00000226757		4	-0.39	5.2	-3.7	0.00098	1	-2.2
ENSG00000188069	OR51F1	11	-0.42	4.8	-3.6	0.00109	1	-2.2
ENSG00000280021		11	-0.42	4.8	-3.6	0.00109	1	-2.2
ENSG00000202296	RNU6-1335P	7	-0.34	4.2	-3.6	0.00115	1	-2.2
ENSG00000222496	RN7SKP200	2	0.44	3.9	3.6	0.00131	1	-2.3
ENSG00000204613	TRIM10	6	-0.31	6.7	-3.5	0.00146	1	-2.3
ENSG00000274747	RN7SL627P	17	-0.42	5.5	-3.5	0.00175	1	-2.4
ENSG00000276088	RN7SL620P	17	-0.42	5.5	-3.5	0.00175	1	-2.4
ENSG00000276532		17	-0.42	5.5	-3.5	0.00175	1	-2.4
ENSG00000278701		17	-0.42	5.5	-3.5	0.00175	1	-2.4
ENSG00000251916	RNU1-61P	6	-0.44	4.3	-3.5	0.00175	1	-2.4
ENSG00000200248	RNA5SP214	6	-0.32	4.2	-3.4	0.00194	1	-2.5
ENSG00000222898	RN7SKP97	8	-0.33	3.5	-3.4	0.00197	1	-2.5
ENSG00000252556	RNU6-256P	11	-0.77	4.8	-3.4	0.00223	1	-2.5
ENSG00000188694	KRTAP24-1	21	-0.27	5.8	-3.4	0.00232	1	-2.5
ENSG00000189326	SPANXN4	X	-0.35	4.6	-3.3	0.00262	1	-2.6
ENSG00000250799	PRODH2	19	-0.24	6	-3.3	0.0027	1	-2.6
ENSG00000238926		X	-0.45	5.2	-3.3	0.00295	1	-2.7
ENSG00000200105	RNU6-251P	12	-0.33	5.4	-3.2	0.00318	1	-2.7
ENSG00000222874	RN7SKP33	20	-0.23	5.5	-3.2	0.00348	1	-2.7
ENSG00000251711	RNU6-632P	10	-0.28	4.9	-3.2	0.00389	1	-2.8
ENSG00000207087	RNU6-242P	2	0.69	5	3.2	0.00389	1	-2.8
ENSG00000223265	RNU6-592P	11	0.69	5	3.2	0.00389	1	-2.8
ENSG00000214107	MAGEB1	X	-0.25	5.3	-3.1	0.00393	1	-2.8
ENSG00000276525		20	0.45	5.1	3.1	0.00408	1	-2.8
ENSG00000218472		6	-0.24	12.8	-3.1	0.00454	1	-2.8
ENSG00000202407		X	-0.5	4.7	-3.1	0.00458	1	-2.8
ENSG00000177627	C12orf54	12	-0.22	4.7	-3.1	0.00494	1	-2.9
ENSG00000147874	HAUS6	9	0.3	7.4	3	0.00513	1	-2.9
ENSG00000227344	HAUS6P1	7	0.3	7.4	3	0.00513	1	-2.9
ENSG00000201988		6	-0.98	4.7	-3	0.00514	1	-2.9
ENSG00000242855	RN7SL496P	7	0.24	3.5	3	0.00575	1	-3
ENSG00000160097	FNDC5	1	-0.2	6.8	-3	0.00576	1	-3
ENSG00000196301	HLA-DRB9	6	-0.3	5.3	-3	0.0058	1	-3
ENSG00000276231	PIK3R6	17	0.37	7.8	3	0.00595	1	-3
ENSG00000199332		6	-0.3	5.6	-3	0.00601	1	-3
ENSG00000187569	DPPA3	12	0.32	5.5	3	0.00617	1	-3
ENSG00000164366	CCDC127	5	-0.5	5.6	-2.9	0.00654	1	-3
ENSG00000106689	LHX2	9	-0.2	6.6	-2.9	0.00675	1	-3
ENSG00000172199	OR8U1	11	0.24	4	2.9	0.00707	1	-3
ENSG00000160949	TONSL	8	-0.22	7.3	-2.9	0.00724	1	-3.1
ENSG00000117410	ATP6V0B	1	0.92	4.6	2.9	0.00729	1	-3.1
ENSG00000213366	GSTM2	1	0.27	8.4	2.9	0.00735	1	-3.1
ENSG00000189134	NKAPL	6	0.25	4.9	2.9	0.00741	1	-3.1
ENSG00000167658	EEF2	19	0.42	4.8	2.9	0.00752	1	-3.1
ENSG00000206775	SNORD37	19	0.42	4.8	2.9	0.00752	1	-3.1
ENSG00000200651		17	0.65	6.1	2.9	0.00761	1	-3.1
ENSG00000183130	OR2T11	1	0.26	5.9	2.9	0.00765	1	-3.1
ENSG00000279301		1	0.26	5.9	2.9	0.00765	1	-3.1

Table 2.2. Differential expression results from the case-control analysis

Group 2

ENS_gene	hgnc_symbol	chr	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000204933	CD177P1	19	-0.86	6	-5.5	0.00018	1	-2
ENSG00000204936	CD177	19	-0.86	6	-5.5	0.00018	1	-2
ENSG00000117410	ATP6V0B	1	0.92	4.7	4.3	0.00124	1	-2.6
ENSG00000252556	RNU6-256P	11	-0.77	4.8	-4.3	0.00125	1	-2.6
ENSG00000201988		6	-0.98	4.5	-4.3	0.00126	1	-2.6
ENSG00000251882	RNU6-475P	6	-0.6	5.1	-4.3	0.00127	1	-2.6
ENSG00000276525		20	0.45	5	4.2	0.00134	1	-2.6
ENSG00000164366	CCDC127	5	-0.5	5.5	-4.2	0.00141	1	-2.6
ENSG00000202296	RNU6-1335P	7	-0.34	4.3	-4.1	0.00156	1	-2.6
ENSG00000207087	RNU6-242P	2	0.69	5.2	4.1	0.00158	1	-2.7
ENSG00000223265	RNU6-592P	11	0.69	5.2	4.1	0.00158	1	-2.7
ENSG00000274747	RN7SL627P	17	-0.42	5.4	-4.1	0.00161	1	-2.7
ENSG00000276088	RN7SL620P	17	-0.42	5.4	-4.1	0.00161	1	-2.7
ENSG00000276532		17	-0.42	5.4	-4.1	0.00161	1	-2.7
ENSG00000278701		17	-0.42	5.4	-4.1	0.00161	1	-2.7
ENSG00000069764	PLA2G10	16	-0.35	4.3	-4.1	0.00181	1	-2.7
ENSG00000238379	RNA5SP103	2	-0.59	4.7	-4	0.00209	1	-2.7
ENSG00000200024		4	-0.45	5.1	-3.9	0.00248	1	-2.8
ENSG00000100811	YY1	14	0.57	6.8	3.8	0.00269	1	-2.8
ENSG00000229104	YY1P2	2	0.57	6.8	3.8	0.00269	1	-2.8
ENSG00000200248	RNA5SP214	6	-0.32	4.2	-3.8	0.00289	1	-2.8
ENSG00000224533	TMLHE-AS1	X	-0.33	5.3	-3.7	0.00317	1	-2.9
ENSG00000202407		X	-0.5	4.7	-3.7	0.00333	1	-2.9
ENSG00000154188	ANGPT1	8	0.29	5.6	3.7	0.00342	1	-2.9
ENSG00000276410	HIST1H2BB	6	-0.35	6	-3.7	0.00344	1	-2.9
ENSG00000201594	RNA5SP517	X	-0.32	3.7	-3.7	0.00359	1	-2.9
ENSG00000252996	RNU6-1315P	12	0.36	6	3.6	0.00429	1	-3
ENSG00000187569	DPPA3	12	0.32	5.6	3.6	0.00436	1	-3
ENSG00000207808	MIR27A	19	0.3	8.7	3.5	0.00466	1	-3
ENSG00000207980	MIR23A	19	0.3	8.7	3.5	0.00466	1	-3
ENSG00000267519		19	0.3	8.7	3.5	0.00466	1	-3
ENSG00000276797	MIR24-2	19	0.3	8.7	3.5	0.00466	1	-3
ENSG00000200105	RNU6-251P	12	-0.33	5.4	-3.5	0.00498	1	-3
ENSG00000188069	OR51F1	11	-0.42	4.7	-3.5	0.005	1	-3
ENSG00000280021		11	-0.42	4.7	-3.5	0.005	1	-3
ENSG00000236965	OR52N3P	11	0.34	4.8	3.4	0.00538	1	-3.1
ENSG00000222806	RNA5SP225	6	-0.35	6.3	-3.4	0.00563	1	-3.1
ENSG00000199936		2	-0.4	5.8	-3.4	0.00571	1	-3.1
ENSG00000223271		15	-0.4	5.8	-3.4	0.00571	1	-3.1
ENSG00000222898	RN7SKP97	8	-0.33	3.6	-3.4	0.00582	1	-3.1
ENSG00000276231	PIK3R6	17	0.37	7.6	3.4	0.00597	1	-3.1
ENSG00000183130	OR2T11	1	0.26	5.8	3.4	0.00612	1	-3.1
ENSG00000279301		1	0.26	5.8	3.4	0.00612	1	-3.1
ENSG00000213366	GSTM2	1	0.27	8.2	3.4	0.00618	1	-3.1
ENSG00000199565		1	-0.37	6.7	-3.4	0.00628	1	-3.1
ENSG00000164037	SLC9B1	4	0.45	5.4	3.3	0.00663	1	-3.1
ENSG00000183704	SLC9B1P1	Y	0.45	5.4	3.3	0.00663	1	-3.1
ENSG00000214329	SLC9B1P2	2	0.45	5.4	3.3	0.00663	1	-3.1
ENSG00000227367	SLC9B1P4	22	0.45	5.4	3.3	0.00663	1	-3.1
ENSG00000233867	SLC9B1P3	10	0.45	5.4	3.3	0.00663	1	-3.1

Table 2.3. Differential expression results from the within-group analysis of discordant ASD pairs (group 2)

2.4 Discussion

In this study, we attempted to characterise gene expression in a cohort of ASD concordant and discordant MZ twin pairs. The results from initial QC indicated that the experimental assays had potentially generated unreliable measurements of gene expression, and indeed subsequent differential expression analysis seemed to confirm this. Because of doubts about the quality of the data generated, we are unfortunately not able to draw any conclusions about the transcriptome in ASD.

The three main observations that lead to this conclusion are worth discussing in turn. Firstly, in relation to the raw experimental data, the observed non-normal distribution of probeset intensities. As discussed in the Methods section, inspecting the density for the overall probeset intensities for all of the arrays revealed a second smaller intensity spike. This issue was dealt with in the course of quality control by removing any probesets belonging to the smaller distribution, successfully removing the this signal. One slight concern however, is that no explicit mention of this issue could be found in the literature - perhaps suggesting the data generated is somehow atypical, and that the intensity measurements from the arrays were noisy or otherwise inaccurate.

Secondly, also in relation to the raw experimental data, was the lower than expected similarity between co-twins in pairs, and higher than expected similarity with unrelated individuals in some instances. Global expression profiles are expected to be more highly correlated for related individuals. In this case, just under half of the co-twins were found to be more similar to unrelated individuals than their own co-twins. This could be the result of sample mislabelling or swapping. However, the expression profiles did cluster correctly by sex, making this somewhat unlikely. One potential explanation could be contamination from genomic DNA or another source of RNA - which would account for the overall high levels of correlation that were observed even between unrelated samples.

Thirdly, in relation to the results of differential expression analysis, are the lack of significant hits and unexpectedly high proportion of non-coding RNA transcripts. Typically, a differential expression analysis might be expected to produce inflated

p-values due to unreliable estimates of per gene variance as well as the effect of unmodelled technical effects [34, 37]. The opposite is observed here, with QQ plots showing overall p-value deflation. In order to reliably call a differentially expressed gene, The Microarray Quality Control (MAQC) consortium recommend using a log fold change cutoff of at least 1 (corresponding to a two fold change in abundance) in conjunction with a stringent unadjusted p-value threshold, for example $p < 0.001$ [40]. Others suggest using even higher log fold change values of between 1.5 and 2 [41]. For both case-control and within-group 2 analyses, no single gene identified as having evidence for differential expression meets any of these criteria. As for the constituent genes of the differentially expressed lists, a large number of these are snRNAs, which is surprising since the HuGe array is primarily designed to profile mRNAs, which potentially indicates sample contamination occurred, or that there was some other unknown technical issue with the arrays.

On a related note, an issue that currently receives a great deal of attention in Epigenome-wide association studies (EWAS) is how best to control for the numerous non-biological sources of variation present. These can range from batch effects, for example the date when the samples or arrays were processed and by which lab technician, to technical effects - e.g. placement of the samples on the arrays (commonly the rows and columns on the individual arrays, as well as the individual chip or plate IDs). While there is little in the way of contemporary recommendations on such factors in gene expression studies, it seems likely that many of these same issues could be relevant. We did not spend much time exploring these in the current study due to the small sample size and single experimental batch, and then the subsequent lack of observed genomic inflation (often an indication of the presence of unmodelled technical variation), but future studies may wish to apply some of the more recent recommendations from EWAS, such as explicitly controlling for technical factors such as chip positional effects in the linear modelling strategy, estimating cell composition effects [42], or estimating and controlling for unknown sources of variation using unsupervised methods such as ISVA (independent surrogate variable analysis) [43].

Due to concerns about the reliability of the data generated, the decision was taken not to proceed with any additional analyses or higher level interpretation of the

results. Unfortunately, very little can be gleaned from this experiment as a result. It has not been possible to identify genes or groups of genes with altered expression in ASD cases, as compared to unaffected controls, nor differences in gene expression within ASD discordant twin pairs. Furthermore, we are not able to assess the overall effectiveness of the methodology, in terms of the utility of microarray expression profiling of peripheral blood samples in the investigation of ASD, and whether the experiment is well powered enough to detect the very small differences in expression expected between MZ twins.

While the analysis phase of this study was being carried out, a pilot expression study using RNA-seq, a newer, alternative profiling technology utilising next-generation sequencing to directly measure transcript abundance, was underway. The preliminary results showed promise, with all indications that the expression data was of a better resolution and higher quality, and potentially able to identify much smaller differences in expression. The experiment did go ahead and is described in the next chapter.

References

[1] K. Kim, K. Lee, H. Bang, J. Y. Kim, and J. K. Choi, "Intersection of genetics and epigenetics in monozygotic twin genomes," *Methods*, 2015.

[2] J. T. Bell and T. D. Spector, "A twin approach to unraveling epigenetics," *Trends in Genetics*, vol. 27, no. 3, pp. 116–125, 2011.

[3] J. T. Bell and R. Saffery, "The value of twins in epigenetic epidemiology," *International Journal of Epidemiology*, p. dyr179, 2012.

[4] M. Ketelaar, R. Hofstra, and M. Hayden, "What monozygotic twins discordant for phenotype illustrate about mechanisms influencing genetic forms of neurodegeneration," *Clinical genetics*, vol. 81, no. 4, pp. 325–333, 2012.

[5] J. Van Dongen, P. E. Slagboom, H. H. Draisma, N. G. Martin, and D. I. Boomsma, "The continuing value of twin studies in the omics era," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 640–653, 2012.

[6] E. L. Dempster, R. Pidsley, L. C. Schalkwyk, S. Owens, A. Georgiades, F. Kane, S. Kalidindi, M. Picchioni, E. Kravariti, T. Toulopoulou, *et al.*, "Disease-associated epigenetic changes in monozygotic twins discordant for schizophrenia and bipolar disorder," *Human molecular genetics*, vol. 20, no. 24, pp. 4786–4796, 2011.

[7] N. Matigian, L. Windus, H. Smith, C. Filippich, C. Pantelis, J. McGrath, B. Mowry, and N. Hayward, "Expression profiling in monozygotic twins discordant for bipolar disorder reveals dysregulation of the wnt signalling pathway," *Molecular psychiatry*, vol. 12, no. 9, pp. 815–825, 2007.

[8] F. Borovecki, L. Lovrecic, J. Zhou, H. Jeong, F. Then, H. Rosas, S. Hersch, P. Hogarth, B. Bouzou, R. Jensen, *et al.*, "Genome-wide expression profiling of human blood reveals biomarkers for huntington's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 31, pp. 11023–11028, 2005.

[9] V. W. Hu, B. C. Frank, S. Heine, N. H. Lee, and J. Quackenbush, "Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes," *BMC genomics*, vol. 7, no. 1, p. 118, 2006.

[10] C. Wong, E. L. Meaburn, A. Ronald, T. Price, A. Jeffries, L. Schalkwyk, R. Plomin, and J. Mill, "Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits," *Molecular psychiatry*, 2013.

[11] A. Trouton, F. M. Spinath, and R. Plomin, "Twins early development study (teds): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood," *Twin Research*, vol. 5, no. 5, pp. 444–448, 2002.

[12] B. R. Oliver and R. Plomin, "Twins' early development study (teds): a multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence," *Twin Research and Human Genetics*, vol. 10, no. 01, pp. 96–105, 2007.

[13] C. Haworth, O. S. Davis, and R. Plomin, "Twins early development study (teds): a genetically sensitive investigation of cognitive and behavioral development from childhood to young adulthood," *Twin Research and Human Genetics*, vol. 1, no. 1, pp. 1–9, 2012.

[14] A. P. Association and A. P. A. T. F. on DSM-IV., *Diagnostic and statistical manual of mental disorders: DSM-IV.* Amer Psychiatric Pub Inc, 1994.

[15] F. J. Scott, S. Baron-Cohen, P. Bolton, and C. Brayne, "The cast (childhood asperger syndrome test) preliminary development of a uk screen for main-

stream primary-school-age children," *Autism*, vol. 6, no. 1, pp. 9–31, 2002.

[16] J. Williams, F. Scott, C. Stott, C. Allison, P. Bolton, S. Baron-Cohen, and C. Brayne, "The cast (childhood asperger syndrome test) test accuracy," *Autism*, vol. 9, no. 1, pp. 45–68, 2005.

[17] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *Journal of autism and developmental disorders*, vol. 24, no. 5, pp. 659–685, 1994.

[18] C. Lord, S. Risi, L. Lambrecht, E. H. Cook Jr, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[19] R. C. Team, "R: A language and environment for statistical computing. vienna, austria; 2014," *URL http://www. R-project. org*, 2015.

[20] B. S. Carvalho and R. A. Irizarry, "A framework for oligonucleotide microarray preprocessing," *Bioinformatics*, vol. 26, no. 19, pp. 2363–2367, 2010.

[21] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[22] B. Carvalho, *pd.hugene.1.0.st.v1 - Platform Design Info for Affymetrix HuGene 1.0 st v1*. R package version 3.14.1.

[23] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, *et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[24] affymetrix, "Dabg on gene level for exon and gene arrays - https://stat.ethz.ch/pipermail/bioconductor/2010-september/035475.html," January 2015.

[25] T. Benaglia, D. Chauveau, D. Hunter, and D. Young, "mixtools: An r package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009.

[26] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, "Biomart–biological queries made easy," *BMC genomics*, vol. 10, no. 1, p. 22, 2009.

[27] WEHI, "http://bioinf.wehi.edu.au/software/gendergenes/index.html," January 2016.

[28] L. Carrel and H. F. Willard, "X-inactivation profile reveals extensive variability in x-linked gene expression in females," *Nature*, vol. 434, no. 7031, pp. 400–404, 2005.

[29] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, *et al.*, "The male-specific region of the human y chromosome is a mosaic of discrete sequence classes," *Nature*, vol. 423, no. 6942, pp. 825–837, 2003.

[30] G. K. Smyth, Y. H. Yang, and T. Speed, "Statistical issues in cdna microarray data analysis," in *Functional Genomics*, pp. 111–136, Springer, 2003.

[31] D. Witten and R. Tibshirani, "A comparison of fold-change and the t-statistic for microarray data analysis," *Technical Report, Stanford University.*, 2007.

[32] H. Yang and G. Churchill, "Estimating p-values in small microarray experiments," *Bioinformatics*, vol. 23, no. 1, pp. 38–43, 2007.

[33] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, Springer, 2005.

[34] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

[35] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.

[36] E. B. Robinson, P. Lichtenstein, H. Anckarsäter, F. Happé, and A. Ronald, "Examining and interpreting the female protective effect against autistic behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. 5258–5262, 2013.

[37] A. A. Fodor, T. L. Tickle, and C. Richardson, "Towards the uniform distribution of null p values on affymetrix microarrays," *Genome Biol*, vol. 8, no. 5, p. R69, 2007.

[38] S. J. Barton, S. R. Crozier, K. A. Lillycrop, K. M. Godfrey, and H. M. Inskip, "Correction of unexpected distributions of p values from analysis of whole genome arrays by rectifying violation of statistical assumptions," *BMC genomics*, vol. 14, no. 1, p. 161, 2013.

[39] A. C. Eklund, L. R. Turner, P. Chen, R. V. Jensen, A. R. Kopf-Sill, Z. Szallasi, *et al.*, "Replacing crna targets with cdna reduces microarray cross-hybridization," *Nature biotechnology*, vol. 24, no. 9, pp. 1071–1073, 2006.

[40] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. De Longueville, E. S. Kawasaki, K. Y. Lee, *et al.*, "The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature biotechnology*, vol. 24, no. 9, pp. 1151–1161, 2006.

[41] A. Brazma and J. Vilo, "Gene expression data analysis," *FEBS letters*, vol. 480, no. 1, pp. 17–24, 2000.

[42] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, "Dna methylation ar-

rays as surrogate measures of cell mixture distribution," *BMC bioinformatics*, vol. 13, no. 1, p. 1, 2012.

[43] A. E. Teschendorff, J. Zhuang, and M. Widschwendter, "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies," *Bioinformatics*, vol. 27, no. 11, pp. 1496–1505, 2011.

3 Characterising the transcriptome in autism spectrum disorder using RNA-seq

3.1 Introduction

RNA-seq is a recently developed method for assessing gene expression using next-generation sequencing. The technology has the potential to revolutionise functional studies of psychiatric traits, by enabling more accurate quantification of low abundance transcripts, as well as alternatively spliced, and non-coding regulatory transcripts. To date, RNA-seq has been used to characterise genome-wide patterns of expression for a range of neurodevelopmental and psychiatric conditions including: major depressive disorder (MDD) [1], schizophrenia [2], bipolar disorder [3], and Alzheimer's disease [4].

Whilst there have been RNA-seq studies that have investigated gene expression during normal development in typically developing controls and attempted to relate these findings to developmental conditions [3, 5], we were unable to find any that directly interrogated the global transcriptome in ASD. In what we believe represents the first study of its kind, we revisit the gene expression profiling study described in Chapter 2, this time employing RNA-seq to profile global patterns of gene expression in ASD. As before, the subjects comprise ASD concordant, ASD discordant, and unaffected concordant control MZ twin pairs from the TEDS

longitudinal birth cohort [6]. Genome-wide gene expression profiles are generated and two main analyses carried out. Firstly, by comparing ASD affected with unaffected, age-matched controls across the sample (between-groups), we aim to identify genes and pathways commonly disrupted in ASD - attributable to genetic and/or environmental factors. Secondly, by performing within-group and within-pairs analyses of the discordant ASD pairs, we also hope to identify those genes and pathways disrupted through non-genetic means (potentially sensitive to the environment), both in common across discordant pairs and family-specific respectively. As well as addressing these primary biological questions, we also hope to provide support for previous epigenetic findings (from the methylation study - see Chapter 2), and also to generate a high-quality gene expression dataset for use in a planned integrative analysis, which will attempt to uncover the functional relationship between altered patterns of expression, disrupted epigenetic regulation, and ASD. Finally, in the course of this work, we also investigate a couple of relevant secondary technical and methodological issues including the cross-platform reproducibility of the expression profiles, and a comparison of a selection of RNA-seq geneset testing methods.

3.2 Methods

3.2.1 Subjects

Subjects were taken from the same cohort of concordant ASD, discordant ASD and age-matched control twin-pairs from TEDS, as previously detailed in Chapter 2. To recap, these included:

- **group 1: 6 ASD concordant pairs - both co-twins in pair with formal ASD diagnosis)**
- **group 2: 6 ASD discordant pairs - one co-twin in pair with a formal ASD diagnosis)**

- **group 6: 11 unaffected, concordant control pairs - both co-twins in pair scored less than or equal to the sample mean in total CAST score**

3.2.2 Pilot study

In order to assess the performance of the proposed platform and fine tune the experimental protocol, prior to the main study, a pilot RNA-seq study using 4 control samples from group 6 (low total CAST - unaffected pairs) was conducted using the Illumina TruSeq Stranded Total RNA protocol and HiSeq 2000 sequencing platform (Illumina, USA). Library preparation was carried out by Matt Arno's lab at the Genomics Centre, King's College London, Waterloo Campus, with sequencing of the libraries conducted by Alka Saxena's lab at the BRC Genomics Core Facility, King's College London, Guy's Hospital. After reviewing the QC metrics and results from the preliminary analysis, performing power calculations using the web application *SCOTTY* [7], and consulting the relevant literature (e.g. [8]), the decision was taken to run the full experiment with the same library kit and on the same platform. Only minor adjustments were made to the library preparation protocol, see the Library preparation section below for more details.

3.2.3 Subject selection

When designing RNA-seq experiments, typically researchers attempt to strike a balance between the desired breadth of coverage - in terms of the number of samples sequenced, and depth of coverage - the number of reads per library. The primary motivating factor here is the high cost associated with sequencing, but there is also the potential to generate an abundance of redundant data from excessive coverage of highly expressed transcripts which it would seem sensible to avoid. So while sequencing to higher read depths does lead to an increase in the accuracy of transcript abundance estimation (thereby increasing power to estimate differentially expressed genes)[9], a read count of 40 M reads per sample is sufficient for detecting the abundance of > 99 % of expressed mRNA [10, 11]. Based on these

prior estimates and the data from the pilot study, for the main experiment the decision was taken to sequence 40 samples on one full flow cell of the HiSeq, yielding an estimated 40 M reads per sample. Subjects were selected from groups 1, 2 and 6, representing ASD concordant, ASD discordant, and unaffected control pairs respectively. After assessing the quality of the total RNA obtained from the blood samples (see RNA isolation below), a total of 17 MZ twin pairs and 6 individuals (40 individuals in total) were available for the study (shown in Figure 3.1).

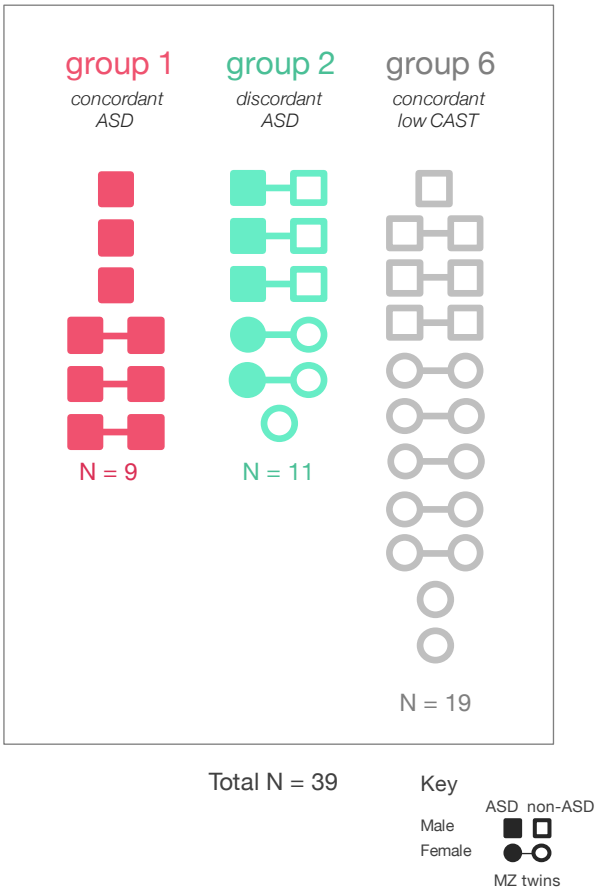

Figure 3.1. The study sample - with group, sex, pair and ASD affected status indicated.

3.2.4 RNA isolation

RNA was isolated from peripheral blood samples, taken from biobanked samples not previously used in the microarray profiling experiment. RNA isolation was performed by Matt Arno's lab. Total RNA was extracted from whole blood samples using the PAXgene blood kit (PreAnalytiX, QIAGEN, Germany). 5 μg total RNA was treated with the Ribo-Zero Globin kit (Ambion, USA) to remove ribosomal and globin RNAs and underwent quality control using the RNA 6000 Nano LabChip and Agilent 2100 BioAnalyzer system (Agilent Technologies, USA). This generates an RNA integrity number (RIN) for each sample, a metric measuring the quality of the starting material [12]. RIN scores range from 1 to 10, with 1 being completely degraded and 10 being intact, and a minimum score of 8 typically recommended to proceed with library preparation [13]. Around 90% of the samples (36/40) were found to be of high quality, with RIN values between 7 and 9 (mean = 8.2). For the remaining samples, RIN values were < 7 (mean = 4.8). The decision was taken to proceed with all but one of one samples : TG36392, from group 1 (ASD concordant), which was found to be severely degraded (RIN = 2.6). Library preparation and sequencing went ahead for the remaining 39 samples.

3.2.5 Library preparation

Libraries of transcripts for each sample were generated. This stage of the experiment was carried out by Matt Arno's lab. The Illumina TruSeq Stranded Total RNA Library Prep Kit v3 (Illumina, USA) was used to prepare RNA libraries following the manufacturers protocol for a low-throughput experiment. First, the purified RNA samples were subjected to thermal fragmentation at 94 degrees for 4 minutes. This was then followed by random hexamer priming. The fragmented, primed RNA then underwent reverse transcription for first strand cDNA synthesis, which was then followed by second strand synthesis, where the RNA template was removed, and a second strand of cDNA synthesised, this time incorporating the strand-marking nucleotide dUTP, to form double stranded cDNA. Following this, end repair, 3' adenylation, and adapter ligation (Illumina Standard and In-

dexed TruSeq adapters) were performed, along with dUTP-marked second strand degradation, to prepare the double stranded cDNA from each of the samples to be multiplexed (pooled) together and run across lanes (to randomize subjects on each lane). Size selection was then performed by electrophoresis, producing fragments with a median length of approximately 326 bp (160 bp inserts). PCR was carried out to enrich for adaptor-bound fragments. The Agilent BioAnalser system was used to assess the distribution of fragment sizes, as a measure of library quality, by inspecting the resulting electropherograms.

3.2.6 Sequencing

The prepared libraries were then sent to the manufacturer Illumina UK for sequencing. Libraries were hybridised to a single flow cell (8 lanes) and sequenced using the HiSeq 2500 (Illumina, USA). This generated 100 bp paired-end reads, yielding an estimated 250 M reads per lane, approximately 50 M reads per sample (with 5 samples per lane). Libraries were arranged in as close to a balanced, randomised design as the available samples would permit, with twin pairs split across the lanes, and each lane containing subjects from each study group (see Appendix - Chapter 3 - Table B1). In order to ensure each lane had the same number of samples and read density, sample TG78042 from group 1 was re-run as a technical replicate, meaning that in total RNA libraries for 40 samples were sequenced (for 39 individuals). The resulting reads were demultiplexed at the BRC Genomics Core Facility, filtering according to the 6 bp index sequences contained within the adaptors and unique to each sample. A total of 80 .fastq files were provided, two files per sample: R1 containing the forward reads and R2 the reverse reads.

3.2.7 Analysis pipeline

Expression profiling by RNA-seq brings with it many novel analytical challenges. While the overall analysis flow is now quite well established (see [14]), consensus has yet to be reached on absolute best practices for each of the analysis stages [8, 15, 16]. A standard RNA-seq differential expression analysis typically proceeds

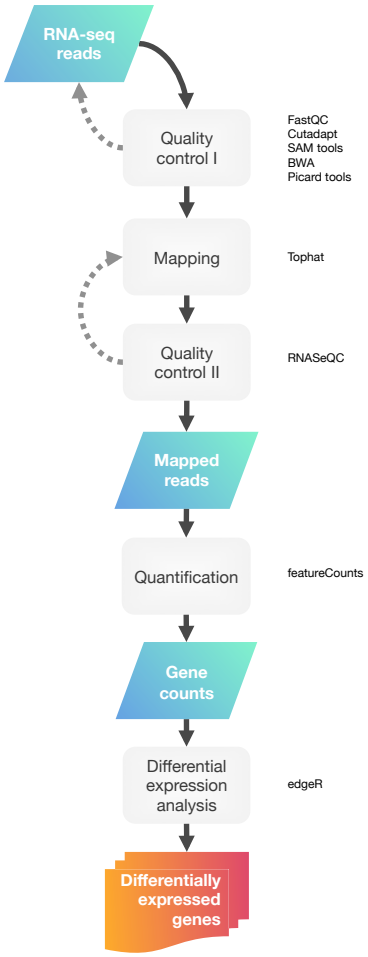

Figure 3.2. The analysis pipeline used to process the RNA-seq data, produce estimates of transcript abundance, and finally test for differential expression between conditions. The software packages used at each stage is indicated.

through the following stages : pre-processing, pre-mapping quality control, mapping, post-mapping quality control, read mapping, quantification and normalization, and finally differential expression analysis. A range of different methods are available for each of these stages, with choice of read mapping and quantification methods arguably having the greatest bearing on the final results produced. For this experiment, a read mapper was chosen based on four main criteria: the underlying method had to have been published, splice-aware (able to map exon-junction spanning reads), with the provided software benchmarked and shown to be computationally fast (as well as supporting parallelisation), and finally, a track

record of being used for large-scale transcriptomic studies. Based on these criteria, the *TopHat* aligner was chosen [17] (see Mapping section below). For later quantification, it was decided that a fast method with parallel support and using minimal statistical modelling would be desirable, to simplify the data being taken forward for later differential expression analysis and data integration. To this end, *featureCounts* from the *subread* package [18] was selected, see Quantification below.

An overview of this analysis pipeline is provided in Fig 3.2, with each step described in the sections below. This scripts are given in Appendix - Chapter 3 - Figure B1, and are also available for download from https://github.com/asaffa/PhD, with software versions, parameters, settings, and additional information also provided.

3.2.8 Quality control I - pre-mapping

The FastQC program [19] was used to produce general sequencing metrics for the .fastq files, to determine the effectiveness of library generation and sequencing, and identify any potential technical biases prior to analysis. The output from pre-mapping QC is given in Appendix - Chapter 3 - Table B2.

3.2.8.1 Number of reads and sequence quality

The number of reads per library ranged between 38 and 78 M, with a mean of approximately 54 M reads per sample. Read depth is a known technical factor in RNA-seq experiments [15], and most analysis flows include some method for normalizing total read count in order to control for this effect, so at this stage no further action was taken. Plots for per-base sequence quality, per-sequence quality, and sequence length distribution were inspected, and were found to be within normal ranges for all samples with no major observable differences between the samples.

3.2.8.2 GC content

GC content was found to range from 46 to 49 % with a median of 47 %, close to the expected 46 % for the human genome [20], confirming that the libraries processed were human RNA, and that contamination from genomic DNA was unlikely to have occurred. For the majority of samples, a twin-peaked distribution for GC count per read was observed, with the higher, narrower peak hypothesised to represent highly expressed transcripts sequenced to an excessive depth. It was reasoned that this might be a characteristic of total RNA libraries derived from whole blood, as this was also observed in the pilot study (conducted at a different sequencing centre). The manufacturer was contacted regarding this issue, but could not offer any explanation. At the time also, no public datasets for whole blood experiments were available for comparison. After some further investigation, it was therefore decided not to take any further action, as there was no evidence for a systematic effect across the lanes (containing randomised samples from each group), with correlation between lane and GC content : $r^2 = 0.25$, $p = 0.12$. It was reasoned too that subsequent normalisation and variance shrinkage should act to reduce any potential bias introduced by the oversampling of particular transcripts - assumed to be the underlying root cause of the GC profile spike.

3.2.8.3 Library composition and sequence diversity

Sequence duplication levels can be used to gain an impression of the diversity of the libraries, and whether over sequencing is likely to have occured. Duplication can either arise from excess amplification of fragments during PCR, or as a result of the same sequence being sequenced multiple times. In this case, the average sequence duplication rate for each library was around 0.21, around the value expected in an RNA-seq experiment [21]. Because there is a degree of uncertainty in determining whether duplication reflects true abundance or PCR artefacts, and because no systematic difference in duplication rates was immediately obvious, it was decided not to remove these.

3.2.8.4 Read removal/trimming

Illumina adapter sequence reads can sometimes make up a small proportion of the libraries due to read through of small fragments. These are flagged as over-represented sequences in *FastQC*. The *cutadapt* tool [22] was used to remove any reads mapping to known Illumina adaptor sequences. Successful removal of these reads was confirmed by re-running *FastQC* and examining the output once again.

Sequencing chemistry generally becomes less reliable with read length, resulting in lower-quality base pair calls towards the ends of reads. *FastQC* generates a boxplot of read quality vs position in order to visualise this. Quality is assessed using the phred score metric $Q : Q = -10 \times \log_{10} P$, where P is the probability of an erroneous base pair call. *cutadapt* was once again used, this time to trim low quality base pair calls, defined as those with phred quality scores <= 20, indicating a 1 in 100 chance of an incorrect base pair call, or 99% calling accuracy.

3.2.8.5 Preliminary mapping and determination of insert sizes

For optimization of the later mapping process, median insert size - defined as fragment length minus the read adaptors (i.e. the actual length of target sequence read) was estimated by mapping subsamples of 1M reads from each sample to the ENSEMBL GRCh37 reference transcriptome using the *BWA* aligner [23]. The *picard tools* program [24] was then used to generate the relevant metrics, and insert sizes were found to range from 169 - 189, with a median of 175 - close to the expected value of 160.

3.2.9 Mapping

The .fastq reads were then mapped to the ENSEMBL GRCh37 reference genome using the *Tophat* [17] software. This is a splice-aware aligner based on the *Bowtie* short read mapper [25], which is rapidly becoming the standard for mapping RNA-seq data, due to its computational efficiency and capability in handling reads span-

ning exon junctions. Binary sequence alignment (.bam) files are generated for each sample.

3.2.10 Quality control II - post mapping

In order to assess the effectiveness of read mapping and the quality of the alignments, *RNA-SeQC* [26] was used to produce summary statistics including total proportion of mapped reads, proportion of reads mapping in pairs, the effectiveness of the stranded protocol, fraction of reads mapping to exonic or intronic regions, and coverage along the length of the transcripts. These metrics were compared to guidelines laid out by GEUVADIS [21]. The overall indications were that the sequencing data generated was of a high quality, with on average 78% of total reads mapping in their forward and reverse read pairs, 47% of these reads originating from exonic regions, a mean coverage of 6 × for less abundant transcripts (bottom 1000) and 732 × for transcripts present at higher levels (top 1000). The output from post-mapping QC is given in Appendix - Chapter 3 - Table B3.

3.2.11 Quantification

A count overlap method was chosen in preference to an isoform abundance estimation method (the method used in *cufflinks* from the *TopHat* suite), as the distributional properties of the former are better understood, allowing greater flexibility with later choice of statistical modelling approach. The *featureCounts* program from the *subread* package [18] was used to summarise and quantify the mapped reads. In order to quantify reads, a counting scheme is used, whereby the number of paired end reads overlapping with annotated genes from the ENSEMBL hg19 transcriptome are totalled. A table of gene counts is output in .csv format, ready for import into *R* for further analysis. The output from post-quantification QC is given in Appendix - Chapter 3 - Table B4.

3.2.12 Filtering

Filtering was performed to remove genes with low counts, as these are potentially unreliable measurements and are unlikely to be of interest for differential expression analysis. Across all of the libraries, reads were found mapping to 46,802 transcripts. A filter requiring abundance >= 1 in at least 3 samples was used, leaving count data for 17,833 transcripts.

3.2.13 Exploratory analysis of inter-individual expression profiles

To assess the quality of the count data, some exploratory analyses were performed. Firstly, concordance between expression profiles within the twin pairs was assessed. As gene expression is heritable, profiles of twins pairs would be expected to be more highly correlated compared to unrelated subjects. For this purpose, counts were normalized by total library size and transformed into logCPM (log counts per million), later reverting back to raw counts for inferential analysis. Correlation was calculated using Pearson's r, using the built-in *cor* function in *R*. A heatmap was produced to visual the results, and mean correlation values within pairs, and between unrelated subjects were recorded.

Next, Euclidean distances between samples were calculated using the complete set of counts, and hierarchical, unsupervised clustering was performed using the *dist* and *hclust* functions in *R*. Two different dendrograms were produced to visualise the results: one in which subject labels were coloured by sex, and the other where subject labels were coloured according to group membership/co-twin status.

Finally, multi-dimensional scaling (MDS) was performed to qualitatively assess the contribution of different biological and technical variables to the overall variance between samples- in this case sex, group, case and lane. MDS is a method which, similar to principal components analysis (PCA), attempts to find a lower dimensional representation of the data, but does so while maintaining distances between objects [27]. In this case, the biological coefficient of variation (BCV) was used as the measure of distance between samples, defined as $\sqrt{\phi_g}$, where ϕ_g is the disper-

sion parameter of the negative binomial distribution (see Differential expression analysis section below). The resulting scatter plots were manually inspected to help highlight any issues.

3.2.14 Differential expression analysis

Count data generated by RNA-seq has particular characteristics that have to be considered before inferential testing can be performed. Firstly, counts are a discrete measurement perhaps best approximated by a negative binomial distribution, in contrast to the continuous, normally distributed intensity measurements from microarrays [16]. Secondly, library size and diversity will vary between sample; the source of the variation can be technical, biological, or stochastic. Thirdly, the number of reads per transcript is proportional to the length of the transcript - longer transcripts will generate more fragments, and hence a greater number of reads.

Normalisation strategies were originally recommended as a means for dealing with some of these issues, for example, the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) measure used by *Cufflinks* in the *Tophat* suite adjusts counts by total number of reads mapped and length of the transcript. However, the current advice is that such normalisation strategies not be used, as they are not robust to the presence of different library sizes and compositions typically found in RNA-seq [28]. More recently developed approaches, such as those implemented in the *bioconductor* packages *DESeq* [29] and *edgeR* use model-based normalization instead of transforming the raw counts. These methods adjust by total library size, using either a scaling factor (median of the ratio of read count over geometric mean across all samples for each gene) or trimmed mean of M-values (TMM - which minimizes logFCs between samples for the majority of genes) respectively. Both of these methods have been empirically shown to be more robust than earlier normalization methods [28].

For differential expression analysis, the *DESeq*,*edgeR* and *limma* methods were first benchmarked using a straight-forward sex analysis comparing expression levels

between males and females (see below) and a "gold standard" set of sex genes - those experimentally confirmed to display sex-specific patterns of expression [30]. Each method's performance was assessed by measuring the sensitivity and specificity for identifying the sex genes as differentially expressed (see Appendix - Chapter 3 - Tables B5, B6, and Figure B2). *edgeR* was selected for the main analysis as it achieved the highest sensitivity, and also because of its similarities with *limma* (previously used in the microarray analysis). *edgeR* also borrows information across genes to improve variance estimates, and a similar linear model-based approach, but one specifically tailored for RNA-seq count data. The method works by calculating a value for common dispersion across all genes, and then using an empirical Bayes approach to shrink the dispersion for each individual gene towards the common trend. Once dispersions have been estimated, differential expression is assessed using a generalized linear model (GLM) likelihood ratio test (LRT). The method proceeds as follows: first negative binomial GLMs for each gene are fitted,

$$Y_{gi} \sim NB(M_i \rho_{gj}, \phi_g) \tag{3.1}$$

where Y_{gi} is the number of counts for gene g in sample i, M_i is the library size, ϕ_g is the dispersion, and ρ_{gj} is the relative abundance of gene g in experimental group j which sample i belongs to. Then, a LRT is used to compare the full model (which includes the covariate of interest) to the null model (where the covariate is excluded), to derive a p-value for evidence of differential expression.

As in the microarray experiment previously (see Chapter 2), two major types of analysis were performed. Firstly, by comparing all cases and controls between-groups (14 ASD affected, 25 unaffected), we intended to identify commonly perturbed genes and pathways associated with ASD that could be genetic or environmental in origin. This analysis will be referred to as "case-control" from here on. The following generalized linear model was used:

$$expression \sim pair + case \tag{3.2}$$

where pair is categorical variable representing twin-pair membership, and case a binary variable denoting case status. Here, we have modelled pair as a fixed effect as $edgeR$, like most negative-binomial distribution-based methods for RNA-seq, assumes statistical independence between all samples and hence does not allow for the fitting of random effects. This same analysis was also repeated using males only, to explicitly control for sex, which we will refer to as "case-control males".

For the second major type of analysis, a within-group comparison of the ASD discordant twins (group 2 : 5 ASD affected, 6 unaffected) was performed, to identify differences in gene expression attributable to the environment. This analysis will be referred to as "group 2". The same model as above was used. This was also repeated using males only, to explicitly control for sex - "group 2 males".

Additionally, for the discordant pairs in group 2, within-pairs (or family-specific) analyses were performed, referred to as "group 2 consensus". For each of the 5 pairs the following model was used:

$$expression \sim case \tag{3.3}$$

Since there are no replicates for the pair unit, this produces only descriptive measures of logFC difference in expression for each gene between case and control. This was used to identify genes with a consistent direction of logFC across all unrelated twin pairs, which could indicate that these genes, as well potentially having their expression modified by environmental factors, could also have a dosage effect giving rise to ASD.

A sex contrast was also performed for benchmarking and quality control purposes, using the following model:

$$expression \sim sex \tag{3.4}$$

To qualitatively assess the overall effectiveness of the statistical modelling approaches used, a number of visualisations were produced: QQ plots to indicate the over-

all distribution of p-values, MA plots to visualize the relationship between logFC and mean count, and volcano plots showing relationship between significance and logFC.

3.2.15 Quantitative trait model

Autistic traits have previously been shown to exist as continuous, normal distributed trait in the general population [31, 32]. An analytical approach that had been planned for the original microarray experiment but later abandoned once it became clear that the data was not usable, was to use total CAST score as a quantitative trait as opposed to the binary case-control outcome in the linear regressions. The idea was revisited in this study.

Before performing differential expression analysis with ASD as a quantitative trait, first descriptive measures for the CAST data were produced. This involved examining the distribution of the data, and exploring the relationship between the scores and other variables such as age. After examining this output it was clear that the CAST score data was not normally distributed and that a data transformation would be required. The Van Der Waerden transformation was used, whereby the data is ranked and then converted to quantiles of the standard normal distribution:

$$VDW = \Phi^{-1}(\frac{R_i}{n+1}) \tag{3.5}$$

where Φ^{-1} is the inverse cumulative distribution function for the normal distribution, R_i is the rank of observation i, and n is the sample size. Further analysis indicated that such a transformation would not be necessary, due to the bimodal nature of the distribution, which meant that for the sample used in this experiment at least, that binary level case-control status would be more appropriate. The observed non-normal CAST score is likely explained by the exclusion of groups 3,4 and 5 from the RNA-seq experiment - these individuals, being above-threshold for one of the triad of traits, are likely to make up the middle (missing) portion of

the distribution. On this basis, it was therefore decided again not to proceed with quantitative trait analysis.

3.2.16 Geneset testing / pathway analysis

Geneset testing was performed for two main purposes. Firstly, small genesets with potential relevance to autism were tested for association, to gauge the success of the experiment in terms of replicating previously identified ASD loci. Secondly, to identify enriched pathways and cellular processes within the differentially expressed genes, to perhaps give some insight as to the underlying processess affected by the dysregulation.

3.2.16.1 Genesets

ASD-associated: To begin with focused testing of ASD-associated genesets was carried out. The first of these sets comprised ASD genes from SFARI (Simons Foundation Autism Research Initiative [33]) gene 2.0 [34], a manually curated reference dataset of ASD genes from published molecular studies including: candidate genes, known common and rare variants, and copy number variants linked to ASD. The second set used was ASD genes from the GWAS catalogue [35] (NHGRI-EBI Catalog of published genome-wide association studies [36]), the entire collection of association findings from GWA studies. The third set consisted of syndromic ASD genes from the Online Mendelian Inheritance in Man (OMIM) [37], a database of genes associated with mendelian disorders. Lastly, the set of all mitochondrial genes discovered across the entire sample was compiled, in order to test for potential enrichment of mitochondrial genes. For benchmarking purposes, a set of sex specific genes from the WEHI bioinformatics group [30] was used. Further information on all of these genesets is given in Table 3.1.

Pathways: Next, geneset testing was performed to test for the enrichment of pathways and cellular processes. These sets were taken from the Molecular Signatures

Gene set	N	Description	Search terms	Link
SFARI	741	Genes implicated in ASD	N/A	https://gene.sfari.org/
GWAS	221	Catalog of published GWAS findings	"autism" (all fields)	https://www.ebi.ac.uk/gwas/
OMIM	388	Genes for mendelian disorders related to ASD	"auti", "[^p]mental","neur" (mim_morbid_description field)	http://www.omim.org/
MT	21	Mitochondrial genes (discovered by seq)	N/A	N/A
XiEgenes	66	Gender specific X chromosome genes	N/A	http://bioinf.wehi.edu.au/software/GenderGenes/
msYgenes	34	Gender specific Y chromosome genes	N/A	http://bioinf.wehi.edu.au/software/GenderGenes/

Table 3.1. ASD relevant genesets used for geneset association testing of the differential expression results

Database (MSigDB) [38] which brings together biochemical pathways (from Kyoto Encyclopedia of Genes and Genomes - KEGG [39] and REACTOME [40]), gene families (from Gene Ontology - GO [41]), and various other sources of annotation, and organises them into various gene set collections of potential biological relevance. Two collections were selected for testing: the C2 curated set - containing 4726 canonical pathways representing metabolic and signalling pathways, and disease gene expression signatures (from microarray studies). The H1 hallmark set was also used - with 50 genesets representing well-defined, biological states or processess, in which genes are known to display coordinated expression patterns.

3.2.16.2 Geneset testing methods

Geneset testing methods can broadly be divided into self-contained (association/ focused) and competitive (enrichment/ battery) methods, depending on the hypothesis being tested. The self-contained null hypothesis is that no genes in the gene set of interest are differentially expressed, whereas the competitive null is that the genes in the set of interest are no more likely to be differentially expressed than its complement (of non- pathway genes outside of the gene set) [42]. Both methods can be further categorised into those using sample randomization or gene randomization to generate the required null distributions for inferential testing, or instead making parametric assumptions [43].

The choice of geneset testing method has been shown to strongly influence the results produced and the conclusions drawn, with different methods potentially producing distinct, non-overlapping sets [44]. However, there is still debate on

which methods should be used to test particular hypotheses in different experimental contexts, and the advice can often be contradictory, see [43] for an in-depth discussion. The same paper demonstrated that both self-contained and competitive methods using sample randomization can produce statistically sound, biological meaningful results, although the use of parametric based methods was cautioned against [43]. This last point is relevant to RNA-seq count data, which is non-normally distributed. Addressing this, Ramatallah et al. used simulated and real data to test geneset methods on RNA-seq data and on this basis recommended the use of self-contained methods based on non-parametric multivariate tests which are found to exert better control over the *Type I* error rate and have increased power over parametric univariate-based methods [45].

For this analysis, it was decided to employ different methods depending on the the characteristics of genesets being tested. In the focused testing of the ASD-associated sets containing candidate genes, since these do not necessarily take part in the same pathways and may or may not be co-expressed, both self-contained and competitive methods would be used, and the consensus taken. And for pathway enrichment testing using the MSigDB pathway sets, which are functionally related and potentially co-expressed sets of genes, only a competitive method would be used - since the aim is to identify pathways with more evidence for enrichment than remaining genes. The *edgeR* package includes a variety of geneset methods designed specifically for use with RNA-seq count data. Here we used the self-contained method *ROAST* [46] and competitive methods *geneSetTest* [47] and *ROMER* [48].

ROAST: The *ROAST* method was selected for self-contained testing of the ASD-associated genesets (SFARI, GWAS, OMIM). *ROAST* uses rotations instead of permutations to generate the null distribution. Rotations can be thought of as a multivariate regression-based analogue of permutations which allow multiple covariates to be randomized, whilst also allowing for correlations [49]. This allows more complex experimental designs with fewer replicates to be used in deriving the null, while avoiding problems with p-value granularity [46]. In simulations and using real data, the method has been found to perform favourably, producing fewer false

positives than a range of other self-contained methods [45]. The following approach is taken: 1. gene level tests are used to obtain moderated t statistics (using the *limma* method), 2. these are converted to z-scores, 3. gene test statistics are then calculated based either on whether a unidirectional or bidirectional hypothesis is being tested (and choice of a summarization method), 4. the data is rotated, 5. and an exact p-value calculated by dividing the number of rotations that produced a statistic as extreme as the observed one by the total number of rotations. Three different summarization methods can be used in combining the t-statistics within the sets, for genes g with z-scores Z_g and optional weights a_g in geneset S, where $A = \sum_{g \in S} |a_g|$ (sum of absolute gene weights of genes in set) the set statistics are computed as:

- T_{mean} : **weighted mean of all the genewise statistics (assumes all genes will be differentially expressed by the same amount)**

$$T_{mean} = \sum_{g \in S} |a_g z_g| / A \tag{3.6}$$

- T_{msq} : **mean of squared genewise statistic (assumes only a small number of genes are differentially expressed)**

$$T_{msq} = \sum_{g \in S} |a_g| z_g^2 / A \tag{3.7}$$

- T_{mean50} : **weighted mean of the top 50% of the genes (S_{50})(half of the genes are differentially expressed)**

$$T_{mean50} = \sum_{g \in S_{50}} |a_g z_g| / A \tag{3.8}$$

- $T_{floormean}$: **floor mean which computes floored z-statistics (similar to mean50, but faster to compute)**

$$T_{floormean} = max(|z_g|, 0.67) \tag{3.9}$$

Since there was no prior hypothesis about the number of differentially expressed genes expected, nor the direction of the effect, a mixed (bidirectional) hypothesis was used, for which the T_{mean50} statistic is generally found to be suitable for. We also decided to implement a further two test statistic summarization methods:

- T_{max} : **maximum of all the other tests - assumes all the scenarios are equally likely**

$$T_{max} = max(T_{mean}, T_{msq}, T_{mean50}, T_{floormean}) \tag{3.10}$$

- T_{artp} : **adaptive rank truncated product (ARTP) - see main text .**

$$T_{artp} = artp(g \in S) \tag{3.11}$$

The adaptive rank truncated product (ARTP) is a general p-value combination method (see Chapter 5) developed by Dudbridge and Koeleman [50], and initially employed in GWAS meta-analysis. ARTP combines the top k most significant gene statistics for a range of truncation points k, selecting the minimum p observed over all points. Yu et al. [51] demonstrated its applicability to gene-based pathway analysis, where it was later shown to outperform a number of alternative methods for both self-contained and competitive tests [52]. The *ROAST* method was modified to include these two additional methods, with the ARTP implementation adapted from the *R* package *ARTP*. The modified *ROAST* function (*TOAST*) can be downloaded from: https://github.com/asaffa/PhD.

geneSetTest: The *geneSetTest* method was used for competitive testing of the ASD sets. The average ranks of the moderated t-statistics in the geneset are calculated, and compared to those in permuted sets in order to obtain a mean overall rank of the geneset. The final p-value is computed using the Wilcoxon test.

ROMER: The *geneSetTest* method assumes genes are independent - which might hold for small sets of candidate genes, but is unlikely to be the case in pathway en-

richment testing. For this reason, an alternative competitive approach was selected for testing of the MSigDB pathway sets. *ROMER* is similar to the popular Gene Set Enrichment Analysis (GSEA) [53] method, but using rotation testing instead of permutations.

3.2.17 Comparison with microarray data

A comparison between the microarray and RNA-seq data was performed with 16 MZ twin pairs and 7 individuals (39 individuals in total) for whom both microarray and RNA-sequencing data were available. Spearman's correlation was used to perform pair-wise comparisons of the samples across the 13,729 genes that were measured on both platforms:

$$\rho_{xy} = r_{R_x R_y} \tag{3.12}$$

where r is Pearson's correlation for the ranks R of the variables x and y. A mean between-platforms sample correlation was calculated, with 95% confidence intervals obtained from Fisher's z transformed values:

$$z = arctanh(r) \tag{3.13}$$

Following this, the levels of correlation and intersection between the differential expression from the different platforms were assessed using the recorded log fold change values for all genes and the significant genes at $p < 0.05$.

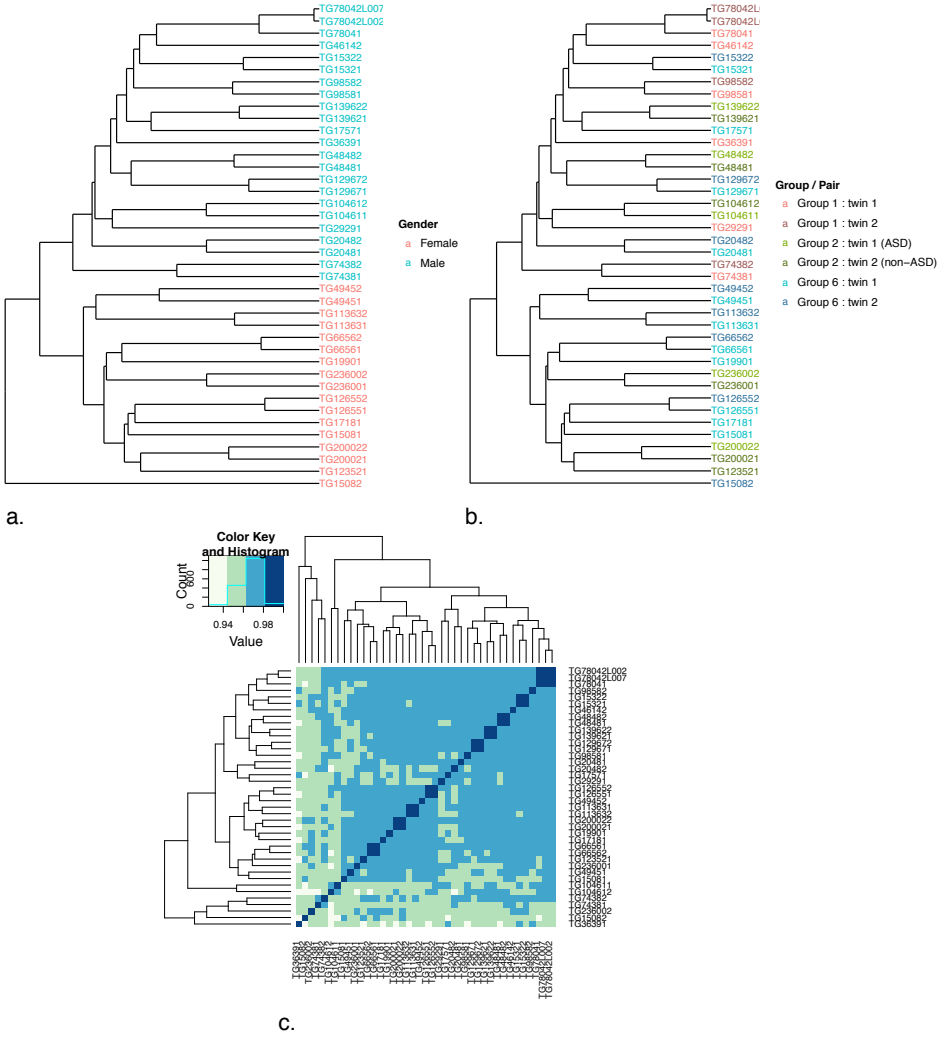

Figure 3.3. a,b. Dendrograms showing sample clustering based on genome-wide expression values. a. labels are coloured by sex, b. labels coloured by group membership and twin pair. Twins clustering together in pairs are concordant for color hue and discordant for shade, and cases of unrelated individuals clustering together are discordant for hue. In this case, all individuals correctly cluster by sex, and all but one pair of twins are found to cluster together as expected. c. Heatmap showing correlations between individuals. High levels of correlation, r values of 0.99 and above, are indicated by the dark blue coloured squares along the diagonal. Twin pairs are found to have an average correlation of 0.98, compared to 0.96 for unrelated subjects.

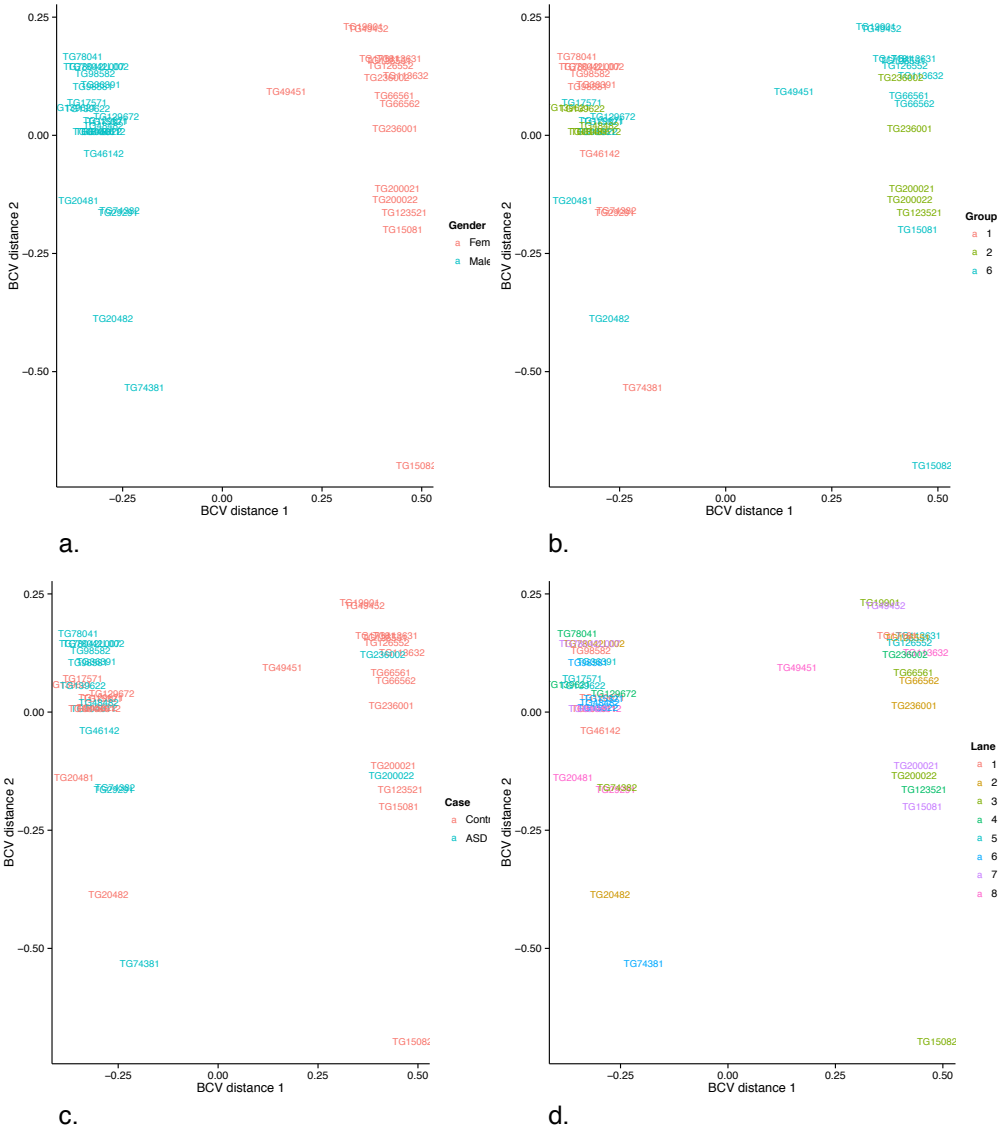

Figure 3.4. MDS plots showing similarity of samples based on BCV distance. a. labels coloured by sex, b. by group membership, c. case status, d. sequencing lane. The majority of the variation between samples projected on to the first two dimensions appears to be mainly accounted for by sex, group and case.

3.3 Results

3.3.1 Exploratory analysis

To begin with, between sample correlation was examined. The median correlation r within pairs was 0.98, compared to a median of 0.96 between unrelated individuals. This is shown in the heatmap in Figure 3.3 c. Next, unsupervised hierarchical clustering was performed using Euclidean distance as the measure. The clustering showed male and female subjects segregating into two distinct clusters with no outliers, and twins were found to cluster together in nodes for 94% (15/16) of the complete MZ twin pairs in the sample. The resulting dendrograms are shown in the top panel of Figures 3.3 a. and b. Finally, MDS was performed to qualitatively examine the contribution of different biological and technical variables to the overall variance between samples. Plots for the first 6 dimensions were inspected, which revealed that sex, case, and group were all contributing to the largest dimension of variance. Lane was not seen to divide the data on any of the first 6 dimensions, and so it was decided on this basis not to include it as a covariate in the linear model. The MDS plots for the first 2 dimensions are given in Figure 3.4. The overall indications from the exploratory analysis were that the count data generated was likely to be of a high quality.

3.3.2 Differential expression analysis

Two main types of analysis were carried out to investigate patterns of gene expression in ASD. Firstly, a between-groups analysis of all affected and unaffected subjects from groups 1,2, and 6 (case-control), which was then repeated excluding all female subjects (case-control males). Secondly, a within-group analysis of the ASD discordant pairs (group 2), which was once again repeated for males only (group 2 males). A within-pairs analysis for each of the ASD discordant pairs was also performed, keeping only genes showing consistent direction of logFC across all the pairs (group 2 consensus).

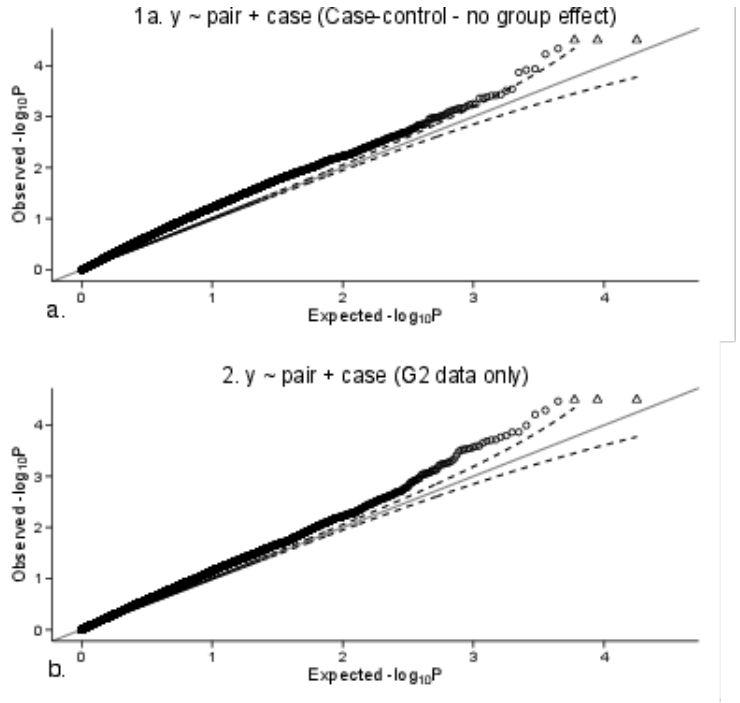

Figure 3.5. QQ plots showing observed vs theoretical distribution of p-values for each of the fitted models

To begin with QQ plots were inspected (plots for case-control and group 2 are shown in Figure 3.5). For all of the models, approximate null distributions were observed, with inflation of p-values nearer the tail. The relative lack of over-inflation of p-values for was taken as an indication that the statistical approach and models used were appropriate, and that there was unlikely to be any large-scale confounding from un-modelled batch effects. Next, MA and volcano plots were produced (case-control and group 2 shown in Figures 3.6, 3.7) to help visualise the proportion of significant differentially expressed genes (at FDR < 0.2). These plots appeared as expected, with only a very small number of highly significant results for each, a roughly equal split between up and down regulated genes, and with signal coming from low as well as high abundance genes. This was taken to indicate that the statistical approach and models used were suitable, and further that they had not appeared to produce any bias towards a particular class of gene (e.g. highly expressed and up-regulated) being identified as differentially expressed.

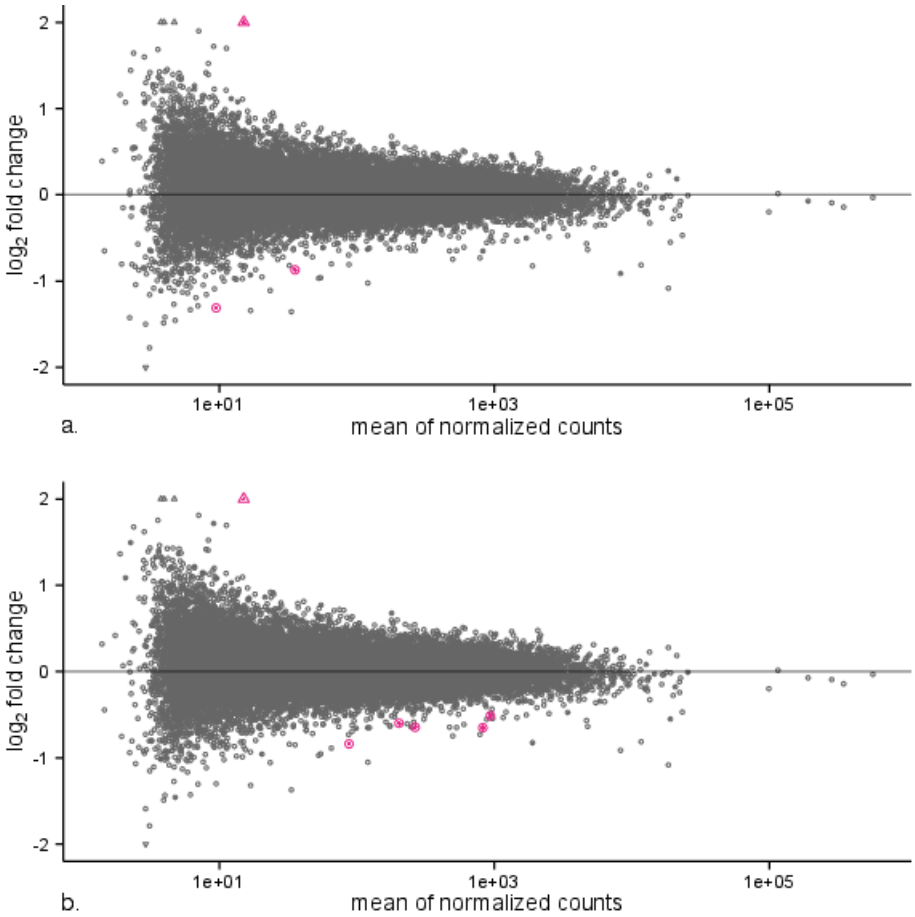

Figure 3.6. MA plots showing the relationship between logFC and mean count for a. case-control and b. group 2 analyses. Significant genes (FDR < 0.2) are indicated by the magenta coloured points, triangles indicate points that lie outside of the plot area. For both of these analyses, only a small number of genes are significant as expected, both up regulated and down regulated, and representing lowly as well as more highly expressed genes.

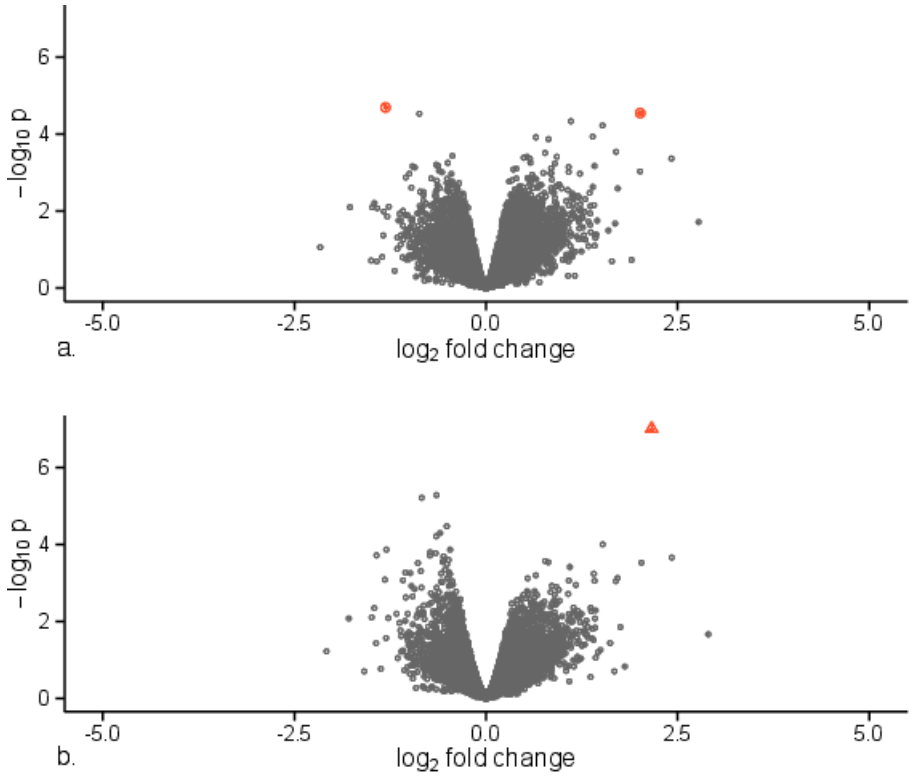

Figure 3.7. Volcano plots showing relationship between significance and logFC for a. case-control and b. group 2 analyses. Significant genes (FDR < 0.2) are indicated by the orange coloured points, triangles indicate points that lie outside of the plot area. For both of these analyses, only a small number of genes are significant, and there appears to be a roughly equal split between up and down regulated genes.

Now turning to the results, the top 50 differentially expressed genes from the case-control analysis are presented in Table 3.2. Overall, while fold changes are modest and the FDR adjusted p-values do not indicate any highly significant findings, a number of genes do pass the FDR < 0.2 threshold, these are: *DEPDC1B* (logFC = -1.31,FDR = 0.18), *IGHG4* (logFC = 2.01, FDR = 0.18), and *ZNF501* (logFC = -0.87, FDR = 0.18). Further down the list, a number of other genes of potential interest (rationale given in Discussion) are found to be nominally significant ($p < 0.05$): *HSPA8P14* (logFC = 2.42, $p = 0.0004$), *HSPA13* (logFC = -0.5,$p = 0.0006$), *SLC15A2* (logFC = 0.4, $p = 0.0008$). Results from the case-control males analysis are shown in Table 3.3. The log fold changes are higher than in the case-control comparison, and once again a number of genes pass the FDR < 0.2 threshold for significance:

HSPA8P14 (logFC = 5.52, FDR = 0.13), *TSPO2* (logFC = -1.85, FDR = 0.14), *IGHG4* (logFC = 1.8, FDR = 0.15), *IGHG3* (logFC = 1.68, FDR = 0.19), *ZNF501* (logFC = 5.52, FDR = 0.13). A nominally significant gene is also highlighted here: *MT-ND5* (logFC = 1.45, $p = 0.0002$).

Case-control

ensembl_gene_id	hgnc_symbol	chr	logFC	logCPM	LR	PValue	FDR
ENSG00000035499	DEPDC1B	5	-1.31	0.4812	18.1	2.1E-05	0.18
ENSG00000211892	IGHG4	14	2.01	0.956	17.5	2.9E-05	0.18
ENSG00000186446	ZNF501	3	-0.87	2.1658	17.4	0.00003	0.18
ENSG00000232184		1	1.11	1.0045	16.6	4.7E-05	0.21
ENSG00000187534	PRR13P5	19	1.52	0.2855	16.1	0.00006	0.21
ENSG00000236029			1.39	0.3512	14.8	0.00012	0.3
ENSG00000224442		2	0.65	2.9969	14.8	0.00012	0.3
ENSG00000127415	IDUA	4	0.82	2.3612	14.6	0.00014	0.3
ENSG00000138395	CDK15	2	1.7	0.6607	13.1	0.00029	0.49
ENSG00000165914	TTC7B	14	0.77	2.8224	13	0.00031	0.49
ENSG00000143401	ANP32E	1	-0.44	4.796	12.7	0.00037	0.49
ENSG00000101104	PABPC1L	20	0.53	3.6236	12.6	0.00039	0.49
ENSG00000260872			0.92	1.5983	12.6	0.00039	0.49
ENSG00000144589	STK11IP	2	0.49	3.6988	12.5	0.00041	0.49
ENSG00000257539	HSPA8P14	12	2.42	-0.4828	12.4	0.00043	0.49
ENSG00000205930	C21orf62-AS1	21	0.57	2.2037	12.4	0.00044	0.49
ENSG00000214226	C17orf67	17	0.58	2.3178	11.9	0.00056	0.51
ENSG00000155304	HSPA13	21	-0.5	4.9262	11.9	0.00056	0.51
ENSG00000176945	MUC20	3	0.9	0.8937	11.8	0.00058	0.51
ENSG00000260093		8	-0.65	1.8065	11.7	0.00063	0.51
ENSG00000153485	TMEM251	14	-0.62	2.1318	11.6	0.00067	0.51
ENSG00000262652		17	1.42	0.2442	11.5	0.00068	0.51
ENSG00000264575	LINC00526	18	-0.96	0.5939	11.5	0.00069	0.51
ENSG00000257390		12	1.08	0.49	11.4	0.00072	0.51
ENSG00000205544	TMEM256	17	-0.93	1.1805	11.4	0.00074	0.51
ENSG00000234737	KRT18P15	3	0.85	1.1455	11.3	0.00077	0.51
ENSG00000163406	SLC15A2	3	0.4	4.934	11.3	0.00079	0.51
ENSG00000137802	MAPKBP1	15	0.36	4.5608	11.2	0.00083	0.51
ENSG00000179271	GADD45GIP1	19	-0.61	2.8082	11	0.00089	0.51
ENSG00000267436		19	2.01	-0.5395	10.9	0.00094	0.51
ENSG00000239697	TNFSF12	17	1.08	0.8882	10.9	0.00097	0.51
ENSG00000075239	ACAT1	11	-0.57	3.0629	10.9	0.00099	0.51
ENSG00000243179		2	0.75	1.3659	10.8	0.001	0.51
ENSG00000174945	AMZ1	7	0.85	0.7247	10.7	0.001	0.51
ENSG00000242931	RPL7P49	17	0.84	1.1475	10.7	0.001	0.51
ENSG00000260339	HEXA-AS1	15	-1	0.622	10.7	0.0011	0.51
ENSG00000265416		2	1.22	0.4022	10.7	0.0011	0.51
ENSG00000136122	BORA	13	-0.45	3.2173	10.7	0.0011	0.51
ENSG00000239382	ALKBH6	19	0.6	1.8584	10.6	0.0011	0.52
ENSG00000237489	LINC00959	10	0.62	1.641	10.3	0.0013	0.53
ENSG00000207137	SNORD116-13	15	-1.05	0.2263	10.3	0.0013	0.53
ENSG00000185829	ARL17A	17	0.54	4.9967	10.2	0.0014	0.53
ENSG00000005882	PDK2	17	0.57	3.0857	10.2	0.0014	0.53
ENSG00000103260	METRN	16	-0.71	1.0787	10.2	0.0014	0.53
ENSG00000240288	GHRLOS	3	0.4	3.8838	10.2	0.0014	0.53
ENSG00000235162	C12orf75	12	-0.51	3.53	10.1	0.0014	0.53
ENSG00000267274		19	0.65	1.8282	10.1	0.0015	0.53
ENSG00000240695		3	0.86	1.6609	10	0.0015	0.53
ENSG00000137817	PARP6	15	0.34	4.8726	10	0.0016	0.53
ENSG00000273148		20	-0.64	1.7008	9.9	0.0016	0.53

Table 3.2. Differential expression results from the between-groups analysis (case-control)

Case-control males

ensembl_gene_id	hgnc_symbol	chr	logFC	logCPM	LR	PValue	FDR
ENSG00000257539	HSPA8P14	12	5.52	-0.4872	19.7	9.2E-06	0.13
ENSG00000265416		2	1.94	0.518	18.8	1.4E-05	0.13
ENSG00000112212	TSPO2	6	-1.85	0.2339	17.9	2.4E-05	0.14
ENSG00000211892	IGHG4	14	1.8	0.4683	17.3	3.3E-05	0.15
ENSG00000243679		7	1.68	0.6411	16.5	4.9E-05	0.17
ENSG00000211897	IGHG3	14	1.68	1.6038	15.9	6.6E-05	0.19
ENSG00000186446	ZNF501	3	-1.08	2.2266	15.6	7.8E-05	0.19
ENSG00000260872			1.21	1.7129	15.4	8.7E-05	0.19
ENSG00000172508	CARNS1	11	1.19	1.9664	15.1	0.0001	0.2
ENSG00000198786	MT-ND5	MT	1.45	4.4545	14.4	0.00015	0.24
ENSG00000267672		19	1.1	1.3494	14.3	0.00016	0.24
ENSG00000245466		14	1.81	0.4484	14.2	0.00016	0.24
ENSG00000230715		7	2.07	0.0641	14	0.00018	0.25
ENSG00000160392	C19orf47	19	1.05	1.7243	13.6	0.00022	0.28
ENSG00000246363		12	-1.46	1.2862	13.4	0.00026	0.29
ENSG00000149534	MS4A2	11	-1.59	0.9971	13.3	0.00026	0.29
ENSG00000211671			1.39	1.2094	13.3	0.00027	0.29
ENSG00000236029			2	0.1635	13.1	0.0003	0.29
ENSG00000267317		19	1.16	1.0758	12.5	0.0004	0.34
ENSG00000230724	LINC01001	11	1.5	2.6541	12.5	0.0004	0.34
ENSG00000140682	TGFB1I1	16	3.03	-0.6621	12.3	0.00046	0.34
ENSG00000070404	FSTL3	19	-1.01	1.3073	12.2	0.00047	0.34
ENSG00000207642	MIR571	4	1.62	0.4666	12.2	0.00048	0.34
ENSG00000221288	MIR663B	2	2.67	-0.2227	12.2	0.00049	0.34
ENSG00000271734		6	1.68	0.3657	12.1	0.0005	0.34
ENSG00000174547	MRPL11	11	-0.64	3.333	12.1	0.00051	0.34
ENSG00000170485	NPAS2	2	1.38	1.0971	12	0.00054	0.34
ENSG00000211692	TRGJP1	7	-1.82	-0.0611	12	0.00054	0.34
ENSG00000136122	BORA	13	-0.59	3.234	11.9	0.00055	0.34
ENSG00000269890		1	1.78	-0.1184	11.8	0.00058	0.35
ENSG00000204209	DAXX	6	0.72	3.6021	11.8	0.00061	0.35
ENSG00000265713		17	1.03	1.6509	11.7	0.00062	0.35
ENSG00000118113	MMP8	11	-1.21	1.1648	11.6	0.00065	0.35
ENSG00000224610		X	1.27	0.7317	11.4	0.00072	0.37
ENSG00000174945	AMZ1	7	1.1	0.8728	11.4	0.00072	0.37
ENSG00000254325		8	0.94	3.4187	11.4	0.00074	0.37
ENSG00000155755	TMEM237	2	-0.84	1.6707	11.1	0.00086	0.39
ENSG00000273165		2	0.93	1.9375	11.1	0.00087	0.39
ENSG00000130827	PLXNA3	X	0.91	3.3488	11	0.00089	0.39
ENSG00000185829	ARL17A	17	0.76	4.9539	10.9	0.00094	0.39
ENSG00000166171	DPCD	10	-0.78	1.8008	10.9	0.00094	0.39
ENSG00000224638		2	1.98	-0.304	10.9	0.00098	0.39
ENSG00000262652		17	1.57	0.4009	10.9	0.00099	0.39
ENSG00000110583	NAA40	11	0.68	3.4195	10.8	0.00099	0.39
ENSG00000162881	OXER1	2	1.07	2.1428	10.8	0.001	0.39
ENSG00000231047	GCNT1P3	3	0.99	1.4212	10.7	0.0011	0.39
ENSG00000100890	KIAA0391	14	-1.08	1.0129	10.7	0.0011	0.39
ENSG00000237489	LINC00959	10	0.82	1.6785	10.6	0.0011	0.39
ENSG00000260093		8	-0.8	1.7674	10.6	0.0011	0.39
ENSG00000267427			1.16	2.2799	10.5	0.0012	0.39

Table 3.3. Differential expression results from the between-groups, males subjects only analysis (case-control males)

The top 50 differentially expressed genes from the group 2 analysis are given in Table 3.4. The log fold changes once again appear to be modest, but here there

is one significant finding at FDR < 0.2 : *IGHG4* (logFC = 2.16, FDR = 0.0003). A number of other genes also pass this threshold: *EVI2A* (logFC = -0.65, FDR = 0.04), *SNORD15B* (logFC = -0.84, FDR = 0.04), *RGS18* (logFC = -0.51, FDR = 0.15), *LPAR6* (logFC = -0.6, FDR = 0.18), *RPL9* (logFC = -0.65, FDR = 0.18). Some nominally significant genes of potential relevance are also highlighted: *DEPDC1B* (logFC = -1.3, p = 0.0001), *HSPA8P14* (logFC = 2.42, p = 0.0002), *HIST1H3J* (logFC = -0.89, p = 0.0003). The differentially expressed gene list for the group 2 males analysis is shown in Table 3.5. Here, the log fold changes are greater than the group 2 analysis with genes that pass the FDR < 0.2 threshold for significance including: *HSPA8P14* (logFC = 5.75, FDR = 0.05), *MT-ND5* (logFC = 1.46, FDR = 0.05), and *IGHG3* (logFC = 1.68, FDR = 0.10). A number of nominally significant genes of interest are also identified: *HIST1H4C* (logFC = -0.76, p = 0.0001), *HIST1H3J* (logFC = -1.02, p = 0.0004), *HIST1H2BL* (logFC = -0.87, p = 0.0005). Finally, Table 3.6 shows the group 2 consensus results. Potential genes of interest with log fold change in a common direction across the discordant twin pairs include: *HSPA8P14* (mean logFC = 3.65), *IGHV3-66* (mean logFC = 1.41), *DEPDC1B* (mean logFC = -1.25).

Group 2

ensembl_gene_id	hgnc_symbol	chr	logFC	logCPM	LR	PValue	FDR
ENSG00000211892	IGHG4	14	2.16	1.8994	32.1	1.5E-08	0.00027
ENSG00000126860	EVI2A	17	-0.65	5.0969	20.7	5.2E-06	0.03646
ENSG00000207445	SNORD15B	11	-0.84	3.0441	20.4	6.1E-06	0.03646
ENSG00000150681	RGS18	1	-0.51	6.9662	17.2	3.4E-05	0.15122
ENSG00000139679	LPAR6	13	-0.6	4.7714	16.4	5.1E-05	0.18161
ENSG00000163682	RPL9	4	-0.65	6.815	16.1	6.2E-05	0.18321
ENSG00000187534	PRR13P5	19	1.52	0.4577	15.1	0.0001	0.24671
ENSG00000163736	PPBP	4	-0.47	7.4058	14.5	0.00014	0.24671
ENSG00000035499	DEPDC1B	5	-1.3	0.5283	14.5	0.00014	0.24671
ENSG00000145425	RPS3A	4	-0.73	6.8576	14.2	0.00016	0.24671
ENSG00000198339			-0.66	3.727	14.1	0.00017	0.24671
ENSG00000138180	CEP55	10	-1.43	0.2215	13.9	0.00019	0.24671
ENSG00000184825			-0.73	3.5947	13.9	0.00019	0.24671
ENSG00000229117	RPL41	12	-0.55	6.9015	13.8	0.0002	0.24671
ENSG00000257539	HSPA8P14	12	2.42	-0.732	13.7	0.00022	0.24671
ENSG00000156482	RPL30	8	-0.49	8.3408	13.4	0.00025	0.24671
ENSG00000166710	B2M	15	-0.55	11.2918	13.4	0.00026	0.24671
ENSG00000165914	TTC7B	14	0.77	2.61	13.3	0.00027	0.24671
ENSG00000122862	SRGN	10	-0.57	9.362	13.2	0.00029	0.24671
ENSG00000127415	IDUA	4	0.81	2.3314	13.1	0.00029	0.24671
ENSG00000267436		19	2.03	-0.7294	13.1	0.0003	0.24671
ENSG00000197153	HIST1H3J	6	-0.89	3.1683	13	0.0003	0.24671
ENSG00000105708	ZNF14	19	-0.51	4.2661	12.9	0.00033	0.25421
ENSG00000232184		1	1.09	0.678	12.6	0.00039	0.28689
ENSG00000205413	SAMD9	7	-0.53	7.2945	12.3	0.00045	0.32267
ENSG00000186446	ZNF501	3	-0.85	2.1971	12.1	0.00049	0.33037
ENSG00000122026	RPL21	13	-0.49	7.1038	12	0.00053	0.33037
ENSG00000200312	RN7SKP255	14	-1.05	3.3458	12	0.00054	0.33037
ENSG00000100890	KIAA0391	14	-0.99	0.7893	11.9	0.00056	0.33037
ENSG00000163221	S100A12	1	-0.58	5.3353	11.9	0.00057	0.33037
ENSG00000236029			1.41	0.4708	11.8	0.00058	0.33037
ENSG00000156508	EEF1A1	6	-0.47	11.5375	11.8	0.00059	0.33037
ENSG00000224442		2	0.65	2.8985	11.7	0.00064	0.34354
ENSG00000155304	HSPA13	21	-0.5	5.0898	11.4	0.00074	0.36643
ENSG00000002726	AOC1	7	1.72	0.1013	11.4	0.00075	0.36643
ENSG00000185829	ARL17A	17	0.54	5.1594	11.3	0.00076	0.36643
ENSG00000177888	ZBTB41	1	-0.51	5.298	11.3	0.00076	0.36643
ENSG00000262202		17	-1.32	0.0213	11.2	0.00083	0.37324
ENSG00000263934	SNORD3A	17	-1.08	10.6755	11.1	0.00085	0.37324
ENSG00000257390		12	1.08	0.4513	11.1	0.00087	0.37324
ENSG00000262652		17	1.42	0.1638	11.1	0.00088	0.37324
ENSG00000143401	ANP32E	1	-0.44	4.988	11	0.0009	0.37324
ENSG00000138395	CDK15	2	1.69	-0.0285	11	0.0009	0.37324
ENSG00000168242			-0.59	4.4376	11	0.00092	0.37384
ENSG00000075239	ACAT1	11	-0.57	3.2791	10.9	0.00097	0.38459
ENSG00000127920	GNG11	7	-0.47	4.4967	10.7	0.0011	0.42065
ENSG00000134419	RPS15A	16	-0.43	6.1772	10.6	0.0011	0.42065
ENSG00000235316	DUSP8P5	10	1.17	0.2446	10.6	0.0011	0.42065
ENSG00000240695		3	0.86	1.4486	10.5	0.0012	0.42486
ENSG00000200959	SNORA74A	5	-0.97	1.9148	10.5	0.0012	0.42486

Table 3.4. Differential expression results from the within-group discordant ASD pairs analysis (group 2)

Group 2 males

ensembl_gene_id	hgnc_symbol	chr	logFC	logCPM	LR	PValue	FDR
ENSG00000257539	HSPA8P14	12	5.75	-0.7224	21.6	3.4E-06	0.049
ENSG00000198786	MT-ND5	MT	1.46	4.3655	20.6	5.5E-06	0.049
ENSG00000211897	IGHG3	14	1.68	2.5136	18.2	0.00002	0.096
ENSG00000198327			-0.9	4.5502	18.1	2.1E-05	0.096
ENSG00000256316			-1.01	4.1314	17.5	2.8E-05	0.1
ENSG00000265416		2	1.94	0.4929	16.3	5.4E-05	0.162
ENSG00000197061	HIST1H4C	6	-0.76	5.593	15	0.00011	0.26
ENSG00000211892	IGHG4	14	1.82	0.6321	14.8	0.00012	0.26
ENSG00000187534	PRR13P5	19	1.64	0.9873	14.5	0.00014	0.273
ENSG00000221288	MIR663B	2	2.75	0.0758	14.1	0.00017	0.31
ENSG00000172508	CARNS1	11	1.19	1.7947	13.5	0.00024	0.391
ENSG00000112212	TSPO2	6	-1.87	0.0914	13.1	0.00029	0.391
ENSG00000243679		7	1.66	0.7064	13	0.00031	0.391
ENSG00000230724	LINC01001	11	1.5	2.2247	12.9	0.00032	0.391
ENSG00000149534	MS4A2	11	-1.59	0.4446	12.8	0.00034	0.391
ENSG00000140682	TGFB1I1	16	3.06	-0.5897	12.6	0.00038	0.391
ENSG00000197153	HIST1H3J	6	-1.02	3.3004	12.6	0.00039	0.391
ENSG00000185829	ARL17A	17	0.76	5.0882	12.6	0.0004	0.391
ENSG00000259962		16	0.95	2.5704	12.4	0.00044	0.391
ENSG00000186446	ZNF501	3	-1.07	2.3449	12.3	0.00046	0.391
ENSG00000185130	HIST1H2BL	6	-0.87	3.2383	12.3	0.00046	0.391
ENSG00000120318	ARAP3	5	0.89	4.4212	11.7	0.00061	0.44
ENSG00000207642	MIR571	4	1.62	0.3808	11.7	0.00064	0.44
ENSG00000245466		14	1.83	0.0667	11.5	0.00069	0.44
ENSG00000182885	ADGRG3	16	0.68	5.1513	11.4	0.00073	0.44
ENSG00000008516	MMP25	16	0.68	6.3193	11.4	0.00073	0.44
ENSG00000184357	HIST1H1B	6	-0.72	5.4308	11.4	0.00075	0.44
ENSG00000011451	WIZ	19	0.97	2.8284	11.3	0.00078	0.44
ENSG00000162881	OXER1	2	1.07	1.5409	11.2	0.00081	0.44
ENSG00000207445	SNORD15B	11	-0.88	2.7396	11	0.0009	0.44
ENSG00000228434		7	0.87	3.4805	11	0.0009	0.44
ENSG00000177732	SOX12	20	1.62	0.9838	11	0.00092	0.44
ENSG00000198888	MT-ND1	MT	1.37	4.8715	10.9	0.00097	0.44
ENSG00000211692	TRGJP1	7	-1.84	0.0784	10.9	0.00098	0.44
ENSG00000196787	HIST1H2AG	6	-0.62	5.1652	10.9	0.00098	0.44
ENSG00000260872			1.2	1.5786	10.8	0.00099	0.44
ENSG00000270108		14	0.83	2.802	10.8	0.001	0.44
ENSG00000230715		7	2.04	0.092	10.8	0.001	0.44
ENSG00000168242			-0.73	4.5693	10.8	0.001	0.44
ENSG00000160392	C19orf47	19	1.07	1.6329	10.7	0.001	0.44
ENSG00000130827	PLXNA3	X	0.91	3.0318	10.7	0.0011	0.44
ENSG00000197459			-0.63	5.3538	10.7	0.0011	0.44
ENSG00000196532			-0.87	3.7928	10.7	0.0011	0.44
ENSG00000259379		15	1.13	1.6171	10.6	0.0011	0.44
ENSG00000236029			1.99	0.0621	10.5	0.0012	0.44
ENSG00000202354	RNY3	7	-1.05	2.1035	10.5	0.0012	0.44
ENSG00000189337	KAZN	1	0.94	2.0255	10.5	0.0012	0.44
ENSG00000246363		12	-1.52	0.4306	10.4	0.0012	0.44
ENSG00000271734		6	1.69	0.1293	10.4	0.0012	0.44
ENSG00000262292		17	1.18	1.4489	10.4	0.0012	0.44

Table 3.5. Differential expression results from the within-group discordant ASD pairs males only analysis (group 2 males)

Group 2 consensus

ensembl_gene_id	hgnc_symbol	chr	mean_logFC
ENSG00000257539	HSPA8P14	12	3.65
ENSG00000092345	DAZL	3	-2.67
ENSG00000269877		19	2.55
ENSG00000002726	AOC1	7	2.44
ENSG00000267436		19	2.35
ENSG00000182586	LINC00334	21	2.07
ENSG00000141086	CTRL	16	2
ENSG00000256591		11	1.95
ENSG00000174697	LEP	7	-1.88
ENSG00000263642	MIR4802	4	1.81
ENSG00000218713		6	1.81
ENSG00000253305	PCDHGB6	5	-1.8
ENSG00000268947		19	1.73
ENSG00000173088	C10orf131	10	1.72
ENSG00000150556	LYPD6B	2	1.7
ENSG00000242114	MTFP1	22	-1.69
ENSG00000259357		1	1.67
ENSG00000226287	TMEM191A	22	-1.67
ENSG00000262202		17	-1.65
ENSG00000270030		11	-1.61
ENSG00000236029			1.6
ENSG00000177465	ACOT4	14	-1.57
ENSG00000207808	MIR27A	19	-1.56
ENSG00000105173	CCNE1	19	-1.56
ENSG00000166033	HTRA1	10	-1.56
ENSG00000241868	RN7SL434P	3	1.52
ENSG00000112742	TTK	6	-1.51
ENSG00000267479			1.51
ENSG00000268550			1.46
ENSG00000207820	MIR545	X	1.45
ENSG00000267342		17	1.45
ENSG00000246228	CASC8	8	1.42
ENSG00000211972	IGHV3-66	14	1.41
ENSG00000203867	RBM20	10	-1.4
ENSG00000214273	AGGF1P1	4	-1.39
ENSG00000237310		7	1.36
ENSG00000237773		7	-1.36
ENSG00000211710	TRBV4-1	7	-1.34
ENSG00000243607	RPL35AP26	11	1.33
ENSG00000255200		11	1.33
ENSG00000162522	KIAA1522	1	1.32
ENSG00000227684	CROCCP4	1	-1.3
ENSG00000226839	MTND6P11	2	1.28
ENSG00000213411	RBM22P2	13	1.25
ENSG00000035499	DEPDC1B	5	-1.25
ENSG00000163293	NIPAL1	4	1.23
ENSG00000214244	SETP21	5	1.22
ENSG00000232097		7	-1.22
ENSG00000147862	NFIB	9	-1.21
ENSG00000259196	HMBOX1-IT1	8	1.18

Table 3.6. P448 results group 2 consensus

3.3.3 Geneset testing

3.3.3.1 ASD-associated

Focused testing of a number of ASD-associated genesets interest was performed, in order to test for enrichment of previously identified risk genes. The results are presented in Table 3.7. The case-control analysis shows no evidence for enrichment of any of the sets across the different tests used, whereas the case-control males analysis shows weak evidence for enrichment of the OMIM (p = 0.05) and MT sets (p = 0.007), based on the results of the *geneSet* test. The group 2 analysis shows no enrichment, while the group 2 males analysis shows enrichment for the MT set (p = 0.002). Since none of the *ROAST* tests returned significant results, the overall conclusion is that there is no evidence for enrichment of previously identified ASD loci the lists of DE genes from the different contrasts.

3.3.3.2 Geneset statistics

In terms of the performance of the different statistical methods in *ROAST*, T_{max} is seen to perform almost identically to T_{mean50}, with T_{ARTP} appearing to be even more conservative in its estimates. This is also confirmed by the sex analysis which tested for enrichment of known sex specific genes. However, possibly due to lack of adequate resolution because of the small sets being tested, significance values produced are identical, making it difficult to draw any further conclusions.

RNA-seq

Case-control

	Significance (mixed hypothesis)			
	Competitive tests	*Self-contained tests*		
Gene set	**geneSet test (LR)**	**ROAST (Tmean50)**	**ROAST (max)**	**ROAST (ARTP)**
SAFARI	1.72E-01	8.26E-02	8.51E-02	1.47E-01
GWAS	6.55E-01	1.54E-01	1.52E-01	2.81E-01
OMIM	4.33E-01	1.59E-01	1.55E-01	2.42E-01
MT	4.80E-01	4.20E-01	4.20E-01	6.10E-01

Case-control males

	Significance (mixed hypothesis)			
	Competitive tests	*Self-contained tests*		
Gene set	**geneSet test (LR)**	**ROAST (Tmean50)**	**ROAST (max)**	**ROAST (ARTP)**
SAFARI	6.22E-02	1.30E-01	1.29E-01	2.41E-01
GWAS	4.39E-01	1.35E-01	1.32E-01	2.48E-01
OMIM	4.92E-02	1.41E-01	1.36E-01	2.14E-01
MT	6.50E-03	9.95E-02	1.17E-01	1.60E-01

Group 2

	Significance (mixed hypothesis)			
	Competitive tests	*Self-contained tests*		
Gene set	**geneSet test (LR)**	**ROAST (Tmean50)**	**ROAST (max)**	**ROAST (ARTP)**
SAFARI	2.52E-01	1.75E-01	1.78E-01	2.31E-01
GWAS	6.60E-01	2.13E-01	2.16E-01	2.53E-01
OMIM	5.77E-01	2.80E-01	2.81E-01	3.16E-01
MT	3.40E-01	5.80E-01	5.80E-01	6.40E-01

Group 2 males

	Significance (mixed hypothesis)			
	Competitive tests	*Self-contained tests*		
Gene set	**geneSet test (LR)**	**ROAST (Tmean50)**	**ROAST (max)**	**ROAST (ARTP)**
SAFARI	1.04E-01	3.22E-01	3.34E-01	2.74E-01
GWAS	4.04E-01	3.22E-01	3.19E-01	2.17E-01
OMIM	1.19E-01	3.36E-01	3.47E-01	3.34E-01
MT	1.50E-03	2.90E-01	3.10E-01	5.40E-01

Gender

	Significance (mixed hypothesis)			
	Competitive tests	*Self-contained tests*		
Gene set	**geneSet test (LR)**	**ROAST (Tmean50)**	**ROAST (max)**	**ROAST (ARTP)**
WEHI_XiE	2.20E-08	1.00E-04	1.00E-04	1.00E-04
WEHI_msY	1.45E-06	1.00E-04	1.00E-04	1.00E-04

Table 3.7. Results from focused testing of ASD-associated genesets. One competitive test - geneSet test was run, alongside the self-contained method ROMER, for which three different statistics were tested. This was performed for each of the experimental contrasts of interest, as indicated by the headings in the table.

3.3.3.3 Pathways

Next, the MSigDB pathways were tested for enrichment. The results are presented in Tables 3.8, 3.9, 3.10, 3.11. Firstly, for the case-control analysis, a number of potentially relevant pathways are found to be significant ($p < 0.05$) including those related to transcriptional control (*HALLMARK_E2F_TARGETS*), immune system function (*BOHN_PRIMARY_IMMUNODEFICIENCY_SYNDROM_UP, BIOCARTA_HSP27_PATHWAY, KYNG_ENVIRONMENTAL_STRESS_RESPONSE_DN*), *PI3K/AKT* cellular signalling (*BIOCARTA_AKT_PATHWAY, REACTOME_PI_3K_CASCADE*). The case-control males analysis also identifies *PI3K/AKT* cellular signalling as being enriched (*REACTOME_PI3K_EVENTS_IN_ERBB2_SIGNALING, REACTOME_PI3K_EVENTS_IN_ERBB4_SIGNALING*), as well as pathways involved in transcriptional control (*REACTOME_RNA_POL_I_TRANSCRIPTION*).

For the group 2 analysis, immune related pathways are once again identified (*BOHN_PRIMARY_IMMUNODEFICIENCY_SYNDROM_UP*), as well as those involved in transcriptional control (*CROSBY_E2F4_TARGETS*). And finally, for the group 2 males analysis, potentially interesting pathways with evidence of enrichment include those involved in: neuronal development(*MODY_HIPPOCAMPUS_PRENATAL*), neurodegeneration(*REACTOME_AMYLOIDS*), immune system function (*KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRO*), transcriptional control (*REACTOME_RNA_POL_I_TRANSCRIPTION, REACTOME_RNA_POL_I_PROMOTER_OPENING*)

Case-control

MSigDB collection	Pathway	NGenes	Up	Down	Mixed
Hallmark (H)					
	HALLMARK_E2F_TARGETS	196	0.993	0.0096	0.0492
	HALLMARK_MITOTIC_SPINDLE	194	0.108	0.6247	0.063
	HALLMARK_MYC_TARGETS_V1	199	0.997	0.0239	0.0934
	HALLMARK_G2M_CHECKPOINT	191	0.978	0.0332	0.1247
	HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	0.964	0.046	0.1539
	HALLMARK_MYC_TARGETS_V2	58	0.704	0.0887	0.1692
	HALLMARK_MTORC1_SIGNALING	196	0.962	0.021	0.1745
	HALLMARK_NOTCH_SIGNALING	27	0.05	0.7497	0.1975
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	108	0.975	0.031	0.2212
	HALLMARK_ANDROGEN_RESPONSE	86	0.953	0.0127	0.241
	HALLMARK_UV_RESPONSE_DN	113	0.311	0.4048	0.2793
	HALLMARK_WNT_BETA_CATENIN_SIGNALING	36	0.077	0.9746	0.2975
	HALLMARK_PROTEIN_SECRETION	90	0.879	0.0669	0.3152
	HALLMARK_TGF_BETA_SIGNALING	49	0.633	0.3169	0.3925
	HALLMARK_APICAL_JUNCTION	139	0.127	0.8187	0.3973
	HALLMARK_ADIPOGENESIS	177	0.821	0.1281	0.441
	HALLMARK_DNA_REPAIR	145	0.924	0.0988	0.4512
	HALLMARK_APICAL_SURFACE	30	0.218	0.7907	0.4564
	HALLMARK_SPERMATOGENESIS	71	0.738	0.2528	0.5541
	HALLMARK_PI3K_AKT_MTOR_SIGNALING	92	0.58	0.1669	0.5616
Curated (C2)					
	SCHAEFFER_PROSTATE_DEVELOPMENT_AND_CANCER…	9	0.343	0.0198	2.00E-04
	BIOCARTA_AKT_PATHWAY	19	0.249	0.1718	7.00E-04
	BOHN_PRIMARY_IMMUNODEFICIENCY_SYNDROM_UP	45	0.955	8.00E-04	0.001
	NIKOLSKY_BREAST_CANCER_21Q22_AMPLICON	14	0.002	0.9222	0.0019
	GUTIERREZ_WALDENSTROEMS_MACROGLOBULINEMIA_1…	9	0.001	0.9751	0.0022
	WEBER_METHYLATED_HCP_IN_SPERM_UP	8	0.039	0.7233	0.003
	BIOCARTA_HSP27_PATHWAY	12	0.287	0.2598	0.003
	CHIARETTI_T_ALL_RELAPSE_PROGNOSIS	19	0.993	7.00E-04	0.0031
	REACTOME_PI_3K_CASCADE	38	0.126	0.473	0.0034
	REACTOME_TAK1_ACTIVATES_NFKB_...	21	0.194	0.5475	0.0039
	LU_TUMOR_VASCULATURE_DN	4	0.925	0.0012	0.004
	ACEVEDO_LIVER_CANCER_WITH_H3K27ME3_DN	106	0.786	8.00E-04	0.005
	BARRIER_CANCER_RELAPSE_TUMOR_SAMPLE_UP	15	0.994	0.0023	0.005
	CROSBY_E2F4_TARGETS	6	0.994	0.0023	0.005
	MYLLYKANGAS_AMPLIFICATION_HOT_SPOT_8	10	0.007	0.6527	0.005
	KYNG_ENVIRONMENTAL_STRESS_RESPONSE_DN	17	0.144	0.0961	0.005
	GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS..	11	0.479	0.1048	0.0053
	SMID_BREAST_CANCER_LUMINAL_A_DN	16	0.694	0.0355	0.0055
	REACTOME_MEIOTIC_SYNAPSIS	41	0.957	0.0029	0.0056
	ASTON_MAJOR_DEPRESSIVE_DISORDER_UP	36	0.428	0.0966	0.0057

Table 3.8. Pathway enrichment for case-control. Results for the top 20 hallmark and top 20 curated sets are given. Here, we record only the result for the mixed hypothesis, which tests for genes in the set being either up or down regulated.

Case-control males

MSigDB collection	Pathway	NGenes	Up	Down	Mixed
Hallmark (H)					
	HALLMARK_MTORC1_SIGNALING	196	0.834	0.043	0.0697
	HALLMARK_DNA_REPAIR	145	0.888	0.0434	0.1126
	HALLMARK_E2F_TARGETS	196	0.941	0.0313	0.1199
	HALLMARK_MYC_TARGETS_V2	58	0.765	0.0634	0.1304
	HALLMARK_G2M_CHECKPOINT	191	0.914	0.0855	0.1463
	HALLMARK_MYC_TARGETS_V1	199	0.937	0.0464	0.1589
	HALLMARK_NOTCH_SIGNALING	27	0.142	0.6496	0.1658
	HALLMARK_P53_PATHWAY	174	0.352	0.2676	0.1677
	HALLMARK_APICAL_SURFACE	30	0.044	0.8624	0.1864
	HALLMARK_WNT_BETA_CATENIN_SIGNALING	36	0.086	0.9407	0.1994
	HALLMARK_HEME_METABOLISM	190	0.793	0.1118	0.202
	HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	0.97	0.0531	0.2195
	HALLMARK_PI3K_AKT_MTOR_SIGNALING	92	0.471	0.3865	0.2327
	HALLMARK_PROTEIN_SECRETION	90	0.894	0.0841	0.2395
	HALLMARK_APICAL_JUNCTION	139	0.143	0.7707	0.2442
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	108	0.931	0.051	0.2836
	HALLMARK_ALLOGRAFT_REJECTION	171	0.838	0.092	0.2873
	HALLMARK_ADIPOGENESIS	177	0.867	0.061	0.2988
	HALLMARK_MITOTIC_SPINDLE	194	0.389	0.6143	0.3002
	HALLMARK_HYPOXIA	154	0.663	0.2923	0.3614
Curated (C2)					
	NAKAMURA_LUNG_CANCER	3	0.968	2.00E-04	4.00E-04
	TRACEY_RESISTANCE_TO_IFNA2_DN	29	0.927	3.00E-04	0.001
	LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_QTL	16	0.396	0.0776	0.0019
	PENG_GLUTAMINE_DEPRIVATION_UP	32	0.349	0.1691	0.0024
	KIM_GLIS2_TARGETS_DN	6	0.881	9.00E-04	0.0029
	PID_P73PATHWAY	67	0.431	0.1142	0.0029
	MURAKAMI_UV_RESPONSE_1HR_UP	12	0.691	0.0129	0.0032
	CHIBA_RESPONSE_TO_TSA_UP	35	0.066	0.2594	0.0032
	BIOCARTA_DNAFRAGMENT_PATHWAY	10	0.951	0.0028	0.004
	REACTOME_RNA_POL_I_TRANSCRIPTION	47	0.939	0.0055	0.0042
	HASLINGER_B_CLL_WITH_6Q21_DELETION	14	0.811	0.0013	0.0051
	ZIRN_TRETINOIN_RESPONSE_WT1_UP	19	0.511	0.2201	0.0061
	KREPPEL_CD99_TARGETS_DN	8	0.018	0.0297	0.0063
	SYED_ESTRADIOL_RESPONSE	19	0.019	0.6894	0.0063
	REACTOME_PI3K_EVENTS_IN_ERBB4_SIGNALING	32	0.128	0.5176	0.007
	LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_SMALL_VS..	36	0.897	0.0027	0.0071
	FINETTI_BREAST_CANCER_KINOME_RED	14	0.967	0.0014	0.0073
	XU_GH1_EXOGENOUS_TARGETS_UP	47	0.239	0.3667	0.0073
	KYNG_RESPONSE_TO_H2O2_VIA_ERCC6_DN	43	0.591	0.0087	0.0093
	REACTOME_PI3K_EVENTS_IN_ERBB2_SIGNALING	36	0.149	0.4985	0.0094

Table 3.9. Pathway enrichment for case-control males. Results for the top 20 hallmark and top 20 curated sets are given. Here, we record only the result for the mixed hypothesis, which tests for genes in the set being either up or down regulated

Group 2

MSigDB collection	Pathway	NGenes	Up	Down	Mixed
Hallmark (H)					
	HALLMARK_E2F_TARGETS	196	0.964	0.0399	0.0715
	HALLMARK_G2M_CHECKPOINT	191	0.921	0.0809	0.1286
	HALLMARK_MITOTIC_SPINDLE	194	0.193	0.6833	0.1336
	HALLMARK_MTORC1_SIGNALING	196	0.936	0.0346	0.1503
	HALLMARK_MYC_TARGETS_V1	199	0.984	0.0579	0.1901
	HALLMARK_SPERMATOGENESIS	71	0.652	0.1957	0.2505
	HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	0.904	0.1151	0.2722
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	108	0.981	0.0569	0.2989
	HALLMARK_NOTCH_SIGNALING	27	0.145	0.7009	0.3338
	HALLMARK_PROTEIN_SECRETION	90	0.832	0.1149	0.4019
	HALLMARK_MYC_TARGETS_V2	58	0.777	0.1377	0.4113
	HALLMARK_ANDROGEN_RESPONSE	86	0.885	0.115	0.4127
	HALLMARK_KRAS_SIGNALING_UP	143	0.96	0.0172	0.4138
	HALLMARK_PANCREAS_BETA_CELLS	13	0.842	0.1031	0.4249
	HALLMARK_APOPTOSIS	141	0.93	0.0375	0.4337
	HALLMARK_UV_RESPONSE_DN	113	0.393	0.5016	0.4492
	HALLMARK_ADIPOGENESIS	177	0.886	0.1469	0.4511
	HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	49	0.77	0.2645	0.4611
	HALLMARK_MYOGENESIS	122	0.245	0.7155	0.4771
	HALLMARK_DNA_REPAIR	145	0.932	0.1652	0.4909
Curated (C2)					
	REACTOME_SIGNALING_BY_CONSTITUTIVELY_ACTIVE_EGF	16	0.943	0.0027	3.00E-04
	BOHN_PRIMARY_IMMUNODEFICIENCY_SYNDROM_UP	45	0.978	0.004	3.00E-04
	NAKAMURA_LUNG_CANCER	3	0.998	0.0017	0.001
	RHODES_UNDIFFERENTIATED_CANCER	65	0.978	0.0031	0.0011
	CROSBY_E2F4_TARGETS	6	0.988	0.0014	0.0012
	REACTOME_N_GLYCAN_ANTENNAE_ELONGATION	11	0.084	0.1034	0.0012
	BARRIER_CANCER_RELAPSE_TUMOR_SAMPLE_UP	15	0.986	0.0014	0.0021
	NIKOLSKY_BREAST_CANCER_21Q22_AMPLICON	14	0.004	0.8622	0.0021
	SCHAEFFER_PROSTATE_DEVELOPMENT_AND_CANCER_BC	4	0.608	0.0304	0.0023
	LU_TUMOR_VASCULATURE_DN	4	0.963	0.0038	0.003
	OHASHI_AURKA_TARGETS	5	0.579	0.1737	0.0035
	NIKOLSKY_BREAST_CANCER_15Q26_AMPLICON	19	0.131	0.1541	0.0038
	REACTOME_NEF_MEDIATED_DOWNREGULATION_OF_MHC_	10	0.05	0.4585	0.004
	HUTTMANN_B_CLL_POOR_SURVIVAL_DN	58	0.161	0.3485	0.004
	REACTOME_ROLE_OF_DCC_IN_REGULATING_APOPTOSIS	7	0.001	0.7037	0.0042
	SASAKI_ADULT_T_CELL_LEUKEMIA	160	0.992	0.0052	0.0042
	LEE_LIVER_CANCER_SURVIVAL_DN	162	0.95	0.0075	0.0051
	SCHAEFFER_PROSTATE_DEVELOPMENT_AND_CANCER…	9	0.51	0.0052	0.0052
	FOURNIER_ACINAR_DEVELOPMENT_LATE_DN	19	0.841	0.0026	0.0055
	GUTIERREZ_WALDENSTROEMS_MACROGLOBULINEMIA_1…	9	0.003	0.9877	0.0055

Table 3.10. Pathway enrichment for group 2. Results for the top 20 hallmark and top 20 curated sets are given. Here, we record only the result for the mixed hypothesis, which tests for genes in the set being either up or down regulated

Group 2 males

MSigDB collection	Pathway	NGenes	Up	Down	Mixed
Hallmark (H)					
	HALLMARK_MTORC1_SIGNALING	196	0.743	0.1155	0.1507
	HALLMARK_HYPOXIA	154	0.839	0.011	0.23
	HALLMARK_E2F_TARGETS	196	0.835	0.0958	0.2371
	HALLMARK_DNA_REPAIR	145	0.882	0.0995	0.268
	HALLMARK_ALLOGRAFT_REJECTION	171	0.875	0.1067	0.2814
	HALLMARK_MYC_TARGETS_V1	199	0.884	0.0947	0.2869
	HALLMARK_MYC_TARGETS_V2	58	0.9	0.0966	0.2903
	HALLMARK_HEME_METABOLISM	190	0.773	0.1566	0.2915
	HALLMARK_G2M_CHECKPOINT	191	0.786	0.2033	0.3079
	HALLMARK_APICAL_SURFACE	30	0.068	0.8845	0.329
	HALLMARK_P53_PATHWAY	174	0.416	0.1373	0.3438
	HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	0.921	0.1235	0.381
	HALLMARK_PI3K_AKT_MTOR_SIGNALING	92	0.553	0.4937	0.4227
	HALLMARK_NOTCH_SIGNALING	27	0.274	0.5993	0.43
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	108	0.922	0.0956	0.4367
	HALLMARK_COMPLEMENT	163	0.76	0.0626	0.44
	HALLMARK_ADIPOGENESIS	177	0.925	0.0718	0.442
	HALLMARK_APICAL_JUNCTION	139	0.265	0.6755	0.456
	HALLMARK_IL6_JAK_STAT3_SIGNALING	72	0.238	0.6492	0.4592
	HALLMARK_APOPTOSIS	141	0.925	0.1084	0.4621
Curated (C2)					
	SASAKI_ADULT_T_CELL_LEUKEMIA	160	0.858	0.0279	2.00E-04
	ZIRN_TRETINOIN_RESPONSE_WT1_UP	19	0.584	0.2119	0.0011
	SERVITJA_LIVER_HNF1A_TARGETS_DN	74	0.382	0.2709	0.0014
	NIKOLSKY_BREAST_CANCER_17Q11_Q21_AMPLICON	71	0.048	0.9539	0.0016
	MODY_HIPPOCAMPUS_PRENATAL	39	0.905	0.0018	0.0018
	BIOCARTA_P27_PATHWAY	13	0.923	0.0047	0.0018
	CHIBA_RESPONSE_TO_TSA_UP	35	0.039	0.3015	0.0019
	KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PROD…	34	0.813	0.0994	0.0019
	GAVIN_PDE3B_TARGETS	20	0.584	0.001	0.0022
	BIOCARTA_FAS_PATHWAY	30	0.438	0.4007	0.003
	LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_SMALL_VS..	36	0.924	0.0061	0.0035
	BIOCARTA_HER2_PATHWAY	18	0.035	0.8545	0.0039
	WARTERS_IR_RESPONSE_5GY	38	0.011	0.6769	0.0041
	RIZ_ERYTHROID_DIFFERENTIATION_HBZ	22	0.591	0.0754	0.0041
	REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING…	37	0.829	0.0063	0.0043
	REACTOME_AMYLOIDS	33	0.95	0.0105	0.0047
	REACTOME_RNA_POL_I_TRANSCRIPTION	47	0.866	0.0046	0.0048
	REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBL.	21	0.335	0.2508	0.0048
	REACTOME_MEIOTIC_RECOMBINATION	41	0.878	3.00E-04	0.006
	REACTOME_RNA_POL_I_PROMOTER_OPENING	24	0.981	0.0113	0.0061

Table 3.11. Pathway enrichment for group 2 males. Results for the top 20 hallmark and top 20 curated sets are given. Here, we record only the result for the mixed hypothesis, which tests for genes in the set being either up or down regulated

3.3.4 Comparison with microarray data

	Measure of correlation			
	logFC all genes		*logFC DE genes (P < 0.05)*	
Comparison	**Correlation (r)**	**Intersection**	**Correlation (r)**	**Intersection**
case control	0.10	0.63	0.36	0.02
case control males	0.22	0.63	0.24	0.03
group 2	0.10	0.63	0.31	0.02
group 2 males	0.22	0.63	0.27	0.02
gender	0.45	0.63	0.84	0.28

Table 3.12. Correlation between the DE gene lists generated by the RNA-seq and microarray experiments. Overall, for each of the analyses, low values of correlation (~ 0.1 - 0.2) are observed when comparing the logFCs for all 13,729 genes measured on both, with the exception of the sex benchmark (0.45). When comparing logFCs for nominally significant (p < 0.05) DE genes, there is little overlap between the gene lists (~ 0.02 - 0.03),but for the genes that do overlap the correlation is slightly better (~ 0.2 to 0.3), once again with the exception of the sex comparison, which shows overlap of almost a third (0.28), and highly correlated logFCs (0.84).

To assess the concordance between RNA-seq and microarray expression profiles, gene abundances were compared for the 39 samples profiled on both technologies, across 13,729 genes measured on both. The mean pair-wise correlation ρ is 0.63, 95% CI[0.4,0.8] ; much lower than the values for cross-platform correlation commonly reported in the literature, typically in the range of 0.8 to 0.9 [54, 55]. Next, differential expression results were compared - see Table 3.12. The correlation between logFC for the different experimental contrasts is low, with values ranging from 0.1 to 0.22. The sex contrast is included for comparative purposes as there are known to be large differences in gene expression between males and females, and here there does appear to be greater correlation between the datasets at 0.45. For the genes identified as being differentially expressed (p < 0.05))the array data and RNA-seq data show little overlap, with overlapping genes ranging from 0.02 to 0.03. Where there is overlap, the level of correlation between reported logFC is higher than it is for all genes, ranging from 0.24 to 0.36 across the different comparisons. Once again, the sex contrast is included, showing much higher correlation at 0.84, and intersection at 0.28.

3.4 Discussion

3.4.1 Differential expression

The results of differential expression analysis did not reveal any large scale differences in gene expression in ASD compared to non-ASD individuals between-groups (case-control), nor within discordant MZ twins (group 2). While numerous genes were found to be differentially expressed, the evidence was not particularly strong, with only a small number of genes passing the already liberal FDR < 0.2 threshold used. There was overlap between the different experimental contrasts, and the expression of a number of genes and pathways was found to be consistently disrupted, which is what we will focus on here.

DEPDC1B was the most highly significant result in the case-control analysis, with a logFC of -1.31 and a nominal p of 2.1×10^{-5}, and notably in the group 2 consensus analysis was also found to be consistently down-regulated within discordant twin pairs, with a mean family-specific logFC of -1.25. This gene is a member of the *DEP* domain coding family, which is involved in regulating cell growth. While it has not previously been associated with ASD, another family member *DEPDC5* has been robustly associated with epilepsy. Given the frequent co-occurence of epilepsy with ASD [56], *DEPDC1B* could also be relevant to ASD, potentially representing a novel ASD risk gene. However, this finding would need to first be validated using PCR, and ideally replicated in another sample, before this could be stated with confidence.

A number of immune related genes were identified in the lists of differentially expressed gene lists, which were found to be consistently up regulated in ASD cases compared to controls. These included the immunoglobins *IGHG4, IGHG3, IGHV3-66* - antibodies involved in immune system response, and the heatshock protein genes *HSPA8P14, HSPA13* - involved in the cellular response to stress. While none of these are believed to have been previously identified as ASD susceptibility genes, the majority of other global gene expression studies into ASD have also implicated immune system related genes (> 70% of the studies summarised in

Chapter 1 - Table 1). Two results that are particularly relevant here are from a study by Voineagu et al. in post mortem tissue samples which found genes related to immunity to be up regulated in cases compared to controls - consistent with what we found here. Garbett et al. also identified heatshock proteins (*HSPA6, HSPB8*) as being differentially regulated, this time in postmortem brain samples from ASD patients [57].

In the wider context of ASD research, immune system disruption is one of the most consistent findings, with epidemiological studies showing that families with ASD affected individuals have a higher rate of autoimmune disorders [58, 59], serological studies finding evidence of increased numbers of activated B and NK cells [60] and elevated levels of pro-inflammatory cytokines [61] in peripheral blood samples, and postmortem studies finding evidence of microglial activation in the dorsolateral prefrontal cortex [62]. In addition, there is also a potential causal link between established environmental factors like pre-natal viral infection and paternal age with immune dysregulation and inflammation and increasing risk of developmental disorder. In the case of viral infection, animal studies have linked maternal influenza infection to altered brain development and behaviour in mouse models of ASD and schizophrenia [63, 64]. As for paternal age, it has been suggested that a link between increased paternal age and increased incidence of ASD, schizophrenia, and bipolar disorder could be the result of immune dysregulation [65], with increases in pro-inflammatory cytokines IL-1β and IL-6 observed in all three disorders [66, 67]. We believe the accumulated evidence from these diverse studies and the potential link to environmental factors makes a compelling case for prioritising the immune related genes identified here for further study in our discordant pairs as well as potentially other twin cohorts, where we then might begin to investigate links to environmental exposures.

3.4.2 Geneset testing

Focused geneset testing of ASD associated genesets of interest indicated that overall there was no evidence for enrichment of previously identified ASD loci. While geneSet test did suggest enrichment of the OMIM and MT sets in the case-control

males analysis, the results from *ROAST* were not in agreement. It may be the case that geneSet test produced inflated estimates in this instance due to the assumption about gene independence being broken, especially in the case of the MT genes, which would be expected to display some level of co-expression.

Two alternative statistical measures were added to *ROAST* and tested here. T_{max} produced almost identical results to T_{mean50}. This suggests that the existing T_{mean50} statistic is already adequate for testing the scenario where there is no strong prior hypothesis about the number of genes expected to be DE in the set and the direction of those differences. The T_{ARTP} produced even more conservative results than the other methods, and given additional computational complexity of the method, in this setting it does not appear to offer much advantage over the other statistics. Further testing using a variety of real and simulated datasets would be required before more general statements could be made about the utility of the different measures. Furthermore, it would be desirable to implement the ARTP method for *ROMER*, as it could perform differently in the competitive testing scenario. These could be carried out as part of a benchmarking study for appraising the performance of several different geneset methods for RNA-seq data.

Pathway enrichment testing was also performed using collections of genes from MSigDB. In the case-control comparison, a number of potentially relevant pathways were identified. For instance, targets of the transcription factor *E2F*, although not previously associated with ASD, does implicate transcriptional control, which is known to be disrupted in monogenic forms of autism such as in Fragile X (see Chapter 1). *PI3K/AKT* pathways were also identified. These pathways are important for regulating neuronal growth, and in addition to previously being associated with ASD [68], mutations in a constituent member of the pathway, *PTEN*, have been observed in monogenic cases of ASD [69]. Within-group 2 analysis identified pathways involved in neurodevelopment and neurodegeneration pathways, which would be worth investigating further, as these were not identified in the overall case-control comparison. In both case-control and group 2 analyses immune system function pathways were identified, which is a consistent finding from microarray studies into ASD. The accumulated evidence would seem to hint at there being a detectable ASD immune signature in blood. Future studies should be designed in

order to establish the molecular drivers, look at the underlying mechanisms, ascertain whether it is primary or secondary to ASD, and whether such immune markers might have utility for prognosis, diagnosis, or endophenotype classification.

3.4.3 Comparison with microarray data

The concordance between the gene expression profiles produced on each platform was found to be much lower than the estimates published previously. Furthermore, there was little agreement between the genes identified as differentially expressed for the different experimental contrasts, with the exception of sex. From these findings we draw two conclusions. Firstly, since the indications were that the RNA-seq measurements of transcript abundance were of a high quality, we believe the non-concordance of the measurements with the microarray intensity readings confirms the previous suspicions about that experiment having not produced accurate measurements. Secondly, as the sex comparison showed a larger intersection and greater correlation across the platforms (despite the likely issues with array data quality), we conclude that these larger differences in expression are less likely to be overwhelmed by noisy or inaccurate measurements. The opposite is also implied - there would appear to be little margin for error when comparing expression between ASD affected and unaffected individuals, as the fold change differences are likely to be small to begin with.

3.4.4 Limitations

There are a number of drawbacks and limitations with the approach taken here. Firstly, there is the assumption that MZ twins are genetically identical. In fact, MZ twins have been found to display all manner of rare genetic differences, for example point mutations, copy number variants, telomere length, uniparental disomy, mosaicism, chromosomal aneuploidies, and mutations in mitochondrial DNA (see [70] for a discussion of these issues). Unfortunately, without the necessary genetic data on the individuals, it is impossible to rule out the possibility that the observed

phenotypic discordances arise from rare genetic events. There is reasons to be optimistic however, as whole-genome sequencing studies have failed to find many replicable differences between MZ twins [71], and there have been few reports of MZ twins having a point mutation causative for a disease [70]. Copy number variants and chromosomal abnormalities are more common, and since such lesions have been repeatedly implicated in the etiology of ASD, we would ideally screen for these. This is outside of the scope of this project, but could make up part of a future study using the same cohort.

Secondly, there is the use of whole blood samples. Whole blood is a highly heterogenous tissue, which raises the possibility of confounding by cell type composition. This has been discussed at length in the context of epigenetic studies where various methods for adjustment have been suggested (see [72]). The potential impact of heterogeneity on expression studies is not as widely discussed, but it is not inconceivable that it could also have an impact. To address this, blood cell counts could be generated for each sample, and a posthoc analysis conducted to test for the association of DE genes with the different cell proportions. A complete dataset was not available in this case, but could perhaps be generated in the course of conducting a follow-up study.

Thirdly, there is also the question of concordance between gene expression in brain and blood and the reliability of individual differences (as discussed in Chapter 1).The results from the different analyses conducted here contain few previously identified ASD risk genes known to be important in the context of neurodevelopment, which is perhaps a little surprising. One further analysis which we could perhaps conduct as a follow up would be to use the Genotype-Tissue Expression (GTEx) database in order to identify which of our top hits are also highly expressed in the brain, and if any are the targets of known eQTLs which might be linked to ASD. There is also the larger issue of whether RNA-seq profiling of whole blood represents an effective way of identifying genes that are likely to be informative about the pathomechanisms of ASD. This is an issue which we did not address here, but would be worth investigating further, as it may be the case that whole blood is primarily useful for investigating changes in gene expression that are secondary to the disorder (to be used for diagnostic or classification purposes), in which case

microarrays might already be adequate.

Finally, it is worth noting that the samples were collected from adolescents with ASD. Since ASD is a disorder of development, it is not clear whether the molecular differences identified reflect those present in early development, or whether they are primary or secondary to onset. Relating this to the previous point about a potential blood-based immune signature for ASD, ideally a prospective birth cohort could used with biological samples collected before ASD diagnosis. The use of such a study design could allow questions to be asked about whether immunological (or other) differences are likely to be relevant to ASD pathology, and whether they have any value as predictive or prognostic biomarkers.

3.4.5 Future directions

It is hoped that the gene expression data generated here can serve as a valuable resource for future work. Over the course of this PhD project, we plan to integrate this data with the DNA methylation data obtained on the same TEDS subjects, in order to provide further support for the findings from the individual studies, identify novel associations, and help place identified genes within their functional context. Longer term, but sadly outside of the scope of this project, we may also wish to use this dataset to investigate other transcriptomic and regulatory phenomena. As with most expression profiling studies, here we have taken what could be considered quite a simplistic view of the relationship between gene expression and phenotype, by aiming to identify differences in overall abundance (i.e. mean expression levels) that are relevant to our trait. Another dimension which would be relatively straight-forward to investigate is whether any differences in the dynamic range of gene expression (i.e. the variability) are apparent in ASD. This type of analysis is becoming popular in EWAS, where it has been used to identify loci displaying methylation variability in cancer [73, 74], type I diabetes [75], and depression [76]. While it is not yet known the extent to which differential variability in DNA methylation plays a role in health and disease, part of the appeal of investigating the phenomenon is its potential to account for some measure of phenotypic plasticity - where the range of responses for a particular genotype is dependent

on the environmental conditions (a type of GxE interaction), which could also be reflected by variation at the epigenome. There is then perhaps an equally strong rationale for investigating variability in gene expression, given the functional link between epigenome and transcriptome. This could be particularly interesting in the context of ASD, where as we have discussed, epigenetic and gene regulatory mechanisms are increasingly believed to play an important role.

Alternative isoform expression represents another possible avenue for future investigations. The vast majority of genes (92-94%) undergo alternative splicing [77], which allows multiple transcripts to be generated from a single coding region and thus contributes to transcriptome complexity. There is interest in studying alternative splicing in the context of disease, where it is thought that perhaps as many as 50% of diseases could arise from mutations disrupting alternative splicing mechanisms [78]. Alternative splicing is thought to play a particularly important role in the brain, and indeed disruption to splicing regulation has been consistently implicated in the pathology of neurodegenerative disease like Parkinson's [79]. As for neurodevelopment and ASD, at least two studies have found that mutations in the neuroligins *NLGN3* and *NLGN4*, which have been robustly associated with ASD, result in the expression of alternative isoforms [80, 81], which it is hypothesised leads failure of cell recognition during synapse formation [80]. Another study by Vioneagu et al. identified dysregulated splicing of a number of targets of *A2BP1*, which showed enrichment for proteins involved in cytoskeleton reorganization and synaptogenesis [82]. Somewhat related to isoform expression, finally, we may also wish to look at allele-specific expression, where genes are silenced in parent-of-origin specific manner resulting in the expression of only one of the parental alleles. Such mechanisms have been shown to be relevant to ASD risk, for example at the *CNTNAP2* locus [83], and at the imprinted 15q11-q13 locus, where loss of a paternal imprinting for the entire region, or maternal imprinting of the *UBE3A* gene is causative for Prader-Willi or Angelman syndrome respectively [84].

3.4.6 Conclusion

To our knowledge, this study represents the largest of its kind to systematically investigate gene expression differences in ASD MZ twins using RNA-seq methods. While the results did not reveal any large-scale differences in transcript abundance associated with ASD, a number of potential genes of interest were identified including: *DEPDC1B*, *IGHG4, IGHG3, IGHV3-66*,*HSPA8P14, HSPA13*,*SLC15A2*. Pathway analysis revealed that the identified genes might be converging on common pathways including those involved in transcriptional control, immunity and *PI3K/AKT* signalling. These pathways appeared to be disrupted in the case-control analysis, which compared ASD affected with non-ASD affected individuals, as well as with ASD discordant pairs, indicating that they could be sensitive to both genetic and non-genetic factors.

The integrative work proceeded as planned and is described in Chapter 5. In the next Chapter, we take a brief diversion from gene expression in ASD to address a methodological issue in methylation-wide association studies (MWAS) with implications for data integration.

References

[1] S. Mostafavi, A. Battle, X. Zhu, J. Potash, M. Weissman, J. Shi, K. Beckman, C. Haudenschild, C. McCormick, R. Mei, *et al.*, "Type i interferon signaling genes in recurrent major depression: increased expression detected by whole-blood rna sequencing," *Molecular psychiatry*, vol. 19, no. 12, pp. 1267–1274, 2014.

[2] J. Xu, J. Sun, J. Chen, L. Wang, A. Li, M. Helm, S. L. Dubovsky, S.-A. Bacanu, Z. Zhao, and X. Chen, "Rna-seq analysis implicates dysregulation of the immune system in schizophrenia," *BMC genomics*, vol. 13, no. 8, p. 1, 2012.

[3] M. Lin, E. Pedrosa, A. Shah, A. Hrabovsky, S. Maqbool, D. Zheng, and H. M. Lachman, "Rna-seq of human neurons derived from ips cells reveals candidate long non-coding rnas involved in neurogenesis and neuropsychiatric disorders," *PLoS One*, vol. 6, no. 9, p. e23356, 2011.

[4] N. A. Twine, K. Janitz, M. R. Wilkins, and M. Janitz, "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by alzheimer's disease," *PloS one*, vol. 6, no. 1, p. e16266, 2011.

[5] M. N. Ziats and O. M. Rennert, "Identification of differentially expressed micrornas across the developing human brain," *Molecular psychiatry*, vol. 19, no. 7, pp. 848–852, 2014.

[6] C. Haworth, O. S. Davis, and R. Plomin, "Twins early development study (teds): a genetically sensitive investigation of cognitive and behavioral de-

velopment from childhood to young adulthood," *Twin Research and Human Genetics*, vol. 1, no. 1, pp. 1–9, 2012.

[7] M. A. Busby, C. Stewart, C. A. Miller, K. R. Grzeda, and G. T. Marth, "Scotty: a web tool for designing rna-seq experiments to measure differential gene expression," *Bioinformatics*, vol. 29, no. 5, pp. 656–657, 2013.

[8] B. J. Blencowe, S. Ahmad, and L. J. Lee, "Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes," *Genes & development*, vol. 23, no. 12, pp. 1379–1386, 2009.

[9] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in rna-seq: a matter of depth," *Genome research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[10] Y. Liu, J. F. Ferguson, C. Xue, I. M. Silverman, B. Gregory, M. P. Reilly, and M. Li, "Evaluating the impact of sequencing depth on transcriptome profiling in human adipose," 2013.

[11] R. Lei, K. Ye, Z. Gu, and X. Sun, "Diminishing returns in next-generation sequencing (ngs) transcriptome data," *Gene*, vol. 557, no. 1, pp. 82–87, 2015.

[12] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, S. Lightfoot, W. Menzel, M. Granzow, and T. Ragg, "The rin: an rna integrity number for assigning integrity values to rna measurements," *BMC molecular biology*, vol. 7, no. 1, p. 3, 2006.

[13] "Truseq library prep - http://support.illumina.com/," February 2014.

[14] A. Oshlack, M. D. Robinson, M. D. Young, *et al.*, "From rna-seq reads to differential expression results," *Genome biol*, vol. 11, no. 12, p. 220, 2010.

[15] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments," *BMC bioinformatics*, vol. 11, no. 1, p. 94, 2010.

[16] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from rna-seq data," *American journal of botany*, vol. 99, no. 2, pp. 248–256, 2012.

[17] C. Trapnell, L. Pachter, and S. L. Salzberg, "Tophat: discovering splice junctions with rna-seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[18] Y. Liao, G. K. Smyth, and W. Shi, "featurecounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, 2014.

[19] S. Andrews *et al.*, "Fastqc: A quality control tool for high throughput sequence data," *Reference Source*, 2010.

[20] J. Romiguier, V. Ranwez, E. J. Douzery, and N. Galtier, "Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes," *Genome research*, vol. 20, no. 8, pp. 1001–1009, 2010.

[21] P. AC't Hoen, M. R. Friedländer, J. Almlöf, M. Sammeth, I. Pulyakhina, S. Y. Anvar, J. F. Laros, H. P. Buermans, O. Karlberg, M. Brännvall, *et al.*, "Reproducibility of high-throughput mrna and small rna sequencing across laboratories," *Nature biotechnology*, vol. 31, no. 11, pp. 1015–1022, 2013.

[22] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet. journal*, vol. 17, no. 1, pp. pp–10, 2011.

[23] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[24] "Picard tools -http://picard.sourceforge.net/," January 2015.

[25] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature methods*, vol. 9, no. 4, pp. 357–359, 2012.

[26] D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz, "Rna-seqc: Rna-seq met-

rics for quality control and process optimization," *Bioinformatics*, vol. 28, no. 11, pp. 1530–1532, 2012.

[27] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[28] M.-A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, *et al.*, "A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis," *Briefings in bioinformatics*, vol. 14, no. 6, pp. 671–683, 2013.

[29] S. Anders and W. Huber, "Differential expression of rna-seq data at the gene level–the deseq package," *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)*, 2012.

[30] WEHI, "http://bioinf.wehi.edu.au/software/gendergenes/index.html," January 2016.

[31] J. N. Constantino, T. Przybeck, D. Friesen, and R. D. Todd, "Reciprocal social behavior in children with and without pervasive developmental disorders.," *Journal of Developmental & Behavioral Pediatrics*, vol. 21, no. 1, pp. 2–11, 2000.

[32] D. H. Geschwind, "Advances in autism," *Annual review of medicine*, vol. 60, p. 367, 2009.

[33] G. D. Fischbach and C. Lord, "The simons simplex collection: a resource for identification of autism genetic risk factors," *Neuron*, vol. 68, no. 2, pp. 192–195, 2010.

[34] S. N. Basu, R. Kollu, and S. Banerjee-Basu, "Autdb: a gene reference resource for autism research," *Nucleic acids research*, vol. 37, no. suppl 1, pp. D832–D836, 2009.

[35] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, *et al.*, "The nhgri gwas catalog, a curated resource of snp-trait associations," *Nucleic acids research*, vol. 42, no. D1, pp. D1001–D1006, 2014.

[36] "Gwas catalogue - www.ebi.ac.uk/gwas," January 2016.

[37] J. Amberger, C. Bocchini, and A. Hamosh, "A new face and new challenges for online mendelian inheritance in man (omim®)," *Human mutation*, vol. 32, no. 5, pp. 564–567, 2011.

[38] A. Liberzon, "A description of the molecular signatures database (msigdb) web site," in *Stem Cell Transcriptional Networks*, pp. 153–160, Springer, 2014.

[39] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "Kegg for integration and interpretation of large-scale molecular data sets," *Nucleic acids research*, p. gkr988, 2011.

[40] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, *et al.*, "The reactome pathway knowledgebase," *Nucleic acids research*, vol. 42, no. D1, pp. D472–D477, 2014.

[41] G. O. Consortium *et al.*, "The gene ontology (go) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D258–D261, 2004.

[42] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[43] H. Maciejewski, "Gene set analysis methods: statistical models and methodological differences," *Briefings in bioinformatics*, p. bbt002, 2013.

[44] M. C. Wu and X. Lin, "Prior biological knowledge-based approaches for the analysis of genome-wide expression profiles using gene sets and pathways," *Statistical methods in medical research*, vol. 18, no. 6, pp. 577–593, 2009.

[45] Y. Rahmatallah, F. Emmert-Streib, and G. Glazko, "Comparative evaluation of gene set analysis approaches for rna-seq data," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[46] D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader, and G. K. Smyth, "Roast: rotation gene set tests for complex microarray experiments," *Bioinformatics*, vol. 26, no. 17, pp. 2176–2182, 2010.

[47] J. Michaud, K. M. Simpson, R. Escher, K. Buchet-Poyau, T. Beissbarth, C. Carmichael, M. E. Ritchie, F. Schütz, P. Cannon, M. Liu, *et al.*, "Integrative analysis of runx1 downstream pathways and target genes," *BMC genomics*, vol. 9, no. 1, p. 363, 2008.

[48] I. J. Majewski, M. E. Ritchie, B. Phipson, J. Corbin, M. Pakusch, A. Ebert, M. Busslinger, H. Koseki, Y. Hu, G. K. Smyth, *et al.*, "Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells," *Blood*, vol. 116, no. 5, pp. 731–739, 2010.

[49] Ø. Langsrud, "Rotation tests," *Statistics and computing*, vol. 15, no. 1, pp. 53–60, 2005.

[50] F. Dudbridge and B. P. Koeleman, "Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies," *The American Journal of Human Genetics*, vol. 75, no. 3, pp. 424–435, 2004.

[51] K. Yu, Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee, "Pathway analysis by adaptive combination of p-values," *Genetic epidemiology*, vol. 33, no. 8, pp. 700–709, 2009.

[52] M. Evangelou, A. Rendon, W. H. Ouwehand, L. Wernisch, and F. Dudbridge, "Comparison of methods for competitive tests of pathway analysis," *PloS one*, vol. 7, no. 7, p. e41018, 2012.

[53] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting

genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[54] C. Wang, B. Gong, P. R. Bushel, J. Thierry-Mieg, D. Thierry-Mieg, J. Xu, H. Fang, H. Hong, J. Shen, Z. Su, *et al.*, "The concordance between rna-seq and microarray data depends on chemical treatment and transcript abundance," *Nature biotechnology*, vol. 32, no. 9, pp. 926–932, 2014.

[55] S. Zhao, W.-P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of rna-seq and microarray in transcriptome profiling of activated t cells," *PloS one*, vol. 9, no. 1, p. e78644, 2014.

[56] G. Baird, E. Simonoff, A. Pickles, S. Chandler, T. Loucas, D. Meldrum, and T. Charman, "Prevalence of disorders of the autism spectrum in a population cohort of children in south thames: the special needs and autism project (snap)," *The lancet*, vol. 368, no. 9531, pp. 210–215, 2006.

[57] K. Garbett, P. J. Ebert, A. Mitchell, C. Lintas, B. Manzi, K. Mirnics, and A. M. Persico, "Immune transcriptome alterations in the temporal cortex of subjects with autism," *Neurobiology of disease*, vol. 30, no. 3, pp. 303–311, 2008.

[58] H. Ó. Atladóttir, M. G. Pedersen, P. Thorsen, P. B. Mortensen, B. Deleuran, W. W. Eaton, and E. T. Parner, "Association of family history of autoimmune diseases and autism spectrum disorders," *Pediatrics*, vol. 124, no. 2, pp. 687–694, 2009.

[59] A. Keil, J. L. Daniels, U. Forssen, C. Hultman, S. Cnattingius, K. C. Söderberg, M. Feychting, and P. Sparen, "Parental autoimmune diseases associated with autism spectrum disorders in offspring," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 6, p. 805, 2010.

[60] P. Ashwood, B. A. Corbett, A. Kantor, H. Schulman, J. Van de Water, and D. G. Amaral, "In search of cellular immunophenotypes in the blood of children with autism," *PLoS One*, vol. 6, no. 5, p. e19299, 2011.

[61] S. Gupta, S. Aggarwal, B. Rashanravan, and T. Lee, "Th1-and th2-like cytokines in cd4+ and cd8+ t cells in autism," *Journal of neuroimmunology*,

vol. 85, no. 1, pp. 106–109, 1998.

[62] J. T. Morgan, G. Chana, C. A. Pardo, C. Achim, K. Semendeferi, J. Buckwalter, E. Courchesne, and I. P. Everall, "Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism," *Biological psychiatry*, vol. 68, no. 4, pp. 368–376, 2010.

[63] S. H. Fatemi, J. Earle, R. Kanodia, D. Kist, E. S. Emamian, P. H. Patterson, L. Shi, and R. Sidwell, "Prenatal viral infection leads to pyramidal cell atrophy and macrocephaly in adulthood: implications for genesis of autism and schizophrenia," *Cellular and molecular neurobiology*, vol. 22, no. 1, pp. 25–33, 2002.

[64] L. Shi, S. H. Fatemi, R. W. Sidwell, and P. H. Patterson, "Maternal influenza infection causes marked behavioral and pharmacological changes in the offspring," *The Journal of Neuroscience*, vol. 23, no. 1, pp. 297–302, 2003.

[65] M. D. Alter, R. Kharkar, K. E. Ramsey, D. W. Craig, R. D. Melmed, T. A. Grebe, R. C. Bay, S. Ober-Reynolds, J. Kirwan, J. J. Jones, *et al.*, "Autism and increased paternal age related changes in global levels of gene expression regulation," *PloS one*, vol. 6, no. 2, p. e16715, 2011.

[66] H. Jyonouchi, S. Sun, and H. Le, "Proinflammatory and regulatory cytokine production associated with innate and adaptive immune responses in children with autism spectrum disorders and developmental regression," *Journal of neuroimmunology*, vol. 120, no. 1, pp. 170–179, 2001.

[67] R. C. Drexhage, E. M. Knijff, R. C. Padmos, L. v. d. Heul-Nieuwenhuijzen, W. Beumer, M. A. Versnel, and H. A. Drexhage, "The mononuclear phagocyte system and its cytokine inflammatory networks in schizophrenia and bipolar disorder," *Expert Review of Neurotherapeutics*, vol. 10, no. 1, pp. 59–76, 2010.

[68] J. Chen, I. Alberts, and X. Li, "Dysregulation of the igf-i/pi3k/akt/mtor signaling pathway in autism spectrum disorders," *International Journal of Developmental Neuroscience*, vol. 35, pp. 35–41, 2014.

[69] J. Zhou and L. F. Parada, "Pten signaling in autism spectrum disorders," *Current opinion in neurobiology*, vol. 22, no. 5, pp. 873–879, 2012.

[70] T. Kato, K. Iwamoto, C. Kakiuchi, G. Kuratomi, and Y. Okazaki, "Genetic or epigenetic difference causing discordance between monozygotic twins as a clue to molecular basis of mental disorders," *Molecular psychiatry*, vol. 10, no. 7, pp. 622–630, 2005.

[71] J. Van Dongen, P. E. Slagboom, H. H. Draisma, N. G. Martin, and D. I. Boomsma, "The continuing value of twin studies in the omics era," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 640–653, 2012.

[72] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, "Dna methylation arrays as surrogate measures of cell mixture distribution," *BMC bioinformatics*, vol. 13, no. 1, p. 1, 2012.

[73] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, *et al.*, "Increased methylation variation in epigenetic domains across cancer types," *Nature genetics*, vol. 43, no. 8, pp. 768–775, 2011.

[74] A. E. Teschendorff and M. Widschwendter, "Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions," *Bioinformatics*, vol. 28, no. 11, pp. 1487–1494, 2012.

[75] D. S. Paul, A. E. Teschendorff, M. A. Dang, R. Lowe, M. I. Hawa, S. Ecker, H. Beyan, S. Cunningham, A. R. Fouts, A. Ramelius, *et al.*, "Increased dna methylation variability in type 1 diabetes across three immune effector cell types," *Nature communications*, vol. 7, 2016.

[76] A. Cordova-Palomera, M. Fatjo-Vilas, C. Gasto, V. Navarro, M. Krebs, and L. Fananas, "Genome-wide methylation study on depression: differential methylation and variable methylation in monozygotic twins," *Translational psychiatry*, vol. 5, no. 4, p. e557, 2015.

[77] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge, "Alternative isoform regulation in human tissue transcriptomes," *Nature*, vol. 456, no. 7221, pp. 470–476, 2008.

[78] A. J. Matlin, F. Clark, and C. W. Smith, "Understanding alternative splicing: towards a cellular code," *Nature reviews Molecular cell biology*, vol. 6, no. 5, pp. 386–398, 2005.

[79] J. D. Mills and M. Janitz, "Alternative splicing of mrna in the molecular pathology of neurodegenerative diseases," *Neurobiology of aging*, vol. 33, no. 5, pp. 1012–e11, 2012.

[80] S. Jamain, H. Quach, C. Betancur, M. Råstam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, *et al.*, "Mutations of the x-linked genes encoding neuroligins nlgn3 and nlgn4 are associated with autism," *Nature genetics*, vol. 34, no. 1, pp. 27–29, 2003.

[81] Z. Talebizadeh, D. Lam, M. Theodoro, D. Bittel, G. Lushington, and M. Butler, "Novel splice isoforms for nlgn3 and nlgn4 with possible implications in autism," *Journal of medical genetics*, vol. 43, no. 5, pp. e21–e21, 2006.

[82] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[83] D. E. Arking, D. J. Cutler, C. W. Brune, T. M. Teslovich, K. West, M. Ikeda, A. Rea, M. Guy, S. Lin, E. H. Cook, *et al.*, "A common genetic variant in the neurexin superfamily member cntnap2 increases familial risk of autism," *The American Journal of Human Genetics*, vol. 82, no. 1, pp. 160–164, 2008.

[84] B. Horsthemke and J. Wagstaff, "Mechanisms of imprinting of the prader–willi/angelman region," *American journal of medical genetics Part A*, vol. 146, no. 16, pp. 2041–2052, 2008.

4 Estimation of a significance threshold for Epigenome-Wide Association Studies

4.1 Introduction

Epigenetic marks are mitotically heritable chemical modifications to DNA and histone proteins, which act in concert to regulate gene expression across developmental stages and tissues [1]. The most widely studied of these marks is DNA methylation, describing the addition of a methyl group to the 5 carbon of cytosine bases to form 5-methylcytosine (5mC), occurring predominantly in the context of CpG dinucleotides. DNA methylation plays a crucial role in cellular processes such as embryonic development, parental imprinting and X-inactivation. Aberrant methylation patterns have been associated with a number of diseases [2], and variation in methylation between individuals could potentially explain a proportion of phenotypic variance [3]. These observations in particular have led to the popularisation of the epigenome-wide association study (EWAS), a type of epigenome-wide association study (EWAS) which profiles methylomic variation genome-wide in the context of normal development and in disease [4].

The growth of EWAS can be at least partially attributed to the introduction of the Illumina Infinium HumanMethylation450K BeadChip (450K array) (Illumina, San Diego, CA, USA). The Illumina 450K array is a low-cost, high-throughput, plat-

form that interrogates just over 485,500 individual CpG sites across the genome, covering <2% of all known CpG sites, and 99% of RefSeq genes (see [5] for a description of the technology). The platform has been used in investigations into the role of methylomic variation across a range of phenotypes and health conditions including cancers [6, 7], autoimmune disorders [8, 9], psychiatric conditions[10, 11] ,[12], age-related phenotypic changes [13], and environmental exposures [14, 15].

For such studies, as was previously the case with GWAS, the development of standardised experimental design protocols and statistical methods is crucial for ensuring that reported findings are robust, reproducible, and biologically relevant [16, 17, 4]. Whilst there are signs that analytical frameworks are beginning to crystallise for EWAS, there is one particular aspect that has not received much attention, and that is the level of evidence required for a test of association for a differentially methylated CpG site (a differentially methylated position - DMP) to reach genome-wide significance.

Establishment of a significance level for EWAS is complicated by the fact that multiple CpG sites are tested for association simultaneously, and that sites in close proximity can have correlated methylation states (co-methylation) [18, 19, 20]. These are the related problems of multiple testing and dependence, which have previously been addressed in genetic studies, but require careful consideration within the EWAS paradigm due to the specific characteristics of methylation data generated by the 450K.

4.1.1 Characteristics of 450K methylation data

Methylation is itself an inherently variable signal, with states varying by tissue type, cell type, and in certain contexts varying even by haplotype (a phenomenon known as allele specific methylation [21]) . Additionally, there is also uncertainty in the measurements made by methylation assays. Array-based platforms like the 450K must necessarily measure CpG methylation over millions of cells. This means that whilst methylation at the level of an individual cell is a binary measurement, with

a CpG being either methylated or unmethylated, the measurement generated by the array is a continuous, quantitative variable averaged over cells [4]. This means that the experimental data generated in EWAS is subject to levels of error and uncertainty not previously experienced in genetic association studies. On the 450K, the final experimental data generated comes in the form of $Beta$ values ($Beta = \frac{\text{methylated signal}}{(\text{methylated signal+unmethylated signal+100})}$) or M values ($M = log_2(\text{methylated signal+}\frac{1}{(\text{unmethylated signal+1})})$.

Another intriguing property of methylation measurements is that the methylation status of CpGs has been shown to be strongly correlated up to genomic distances of between 1 and 2Kb [18, 19, 20]. This is analogous to linkage disequilibrium (LD - blocks of correlated SNP genotypes) in the genome, with some notable differences. In the case of LD, shared ancestry is responsible for the correlation between SNPs, whereas the precise mechanism generating co-methylation is unknown. It is entirely possible that LD could itself be generating dependency between CpG sites based on their physical proximity. Indeed, a previous study found that some sets of correlated methylated CpG sites appeared to be associated with SNPs in LD blocks [22]. The extent of co-methylation might also depend on functional context [23]. For example, sites located in CpG islands (regions of high CpG density, often found in gene promoters), would be expected to display a high degree of co-methylation, as combined they form a functional unit involved in gene silencing.

The presence of a correlation structure within methylation data has two major implications for downstream analysis. Perhaps most crucially, dependence between measures needs to be taken into account when estimating a significance threshold, for reasons to be discussed. Secondly, and more conveniently, the correlation structure can also be exploited to enable discrete regions to be defined (or discovered, using predefined genomic features).

4.1.2 Levels of analysis

The existence of correlated blocks of methylated CpGs provides a convenient means for grouping together multiple sites. The goal of the analysis then becomes to identify differentially methylated regions (DMRs), which might be preferred over an

individual site-level analysis, aiming to identify differentially methylated positions (DMPs), for a number of reasons. Firstly, regions are considered to be more robust measurement units, since they capture information about the smaller number of independent methylation states that actually exist [4], reducing the impact of measurement uncertainty at the level of individual sites as well as the number of hypotheses being tested. Secondly, regions are potentially a more predictive feature [24], with many reported functionally relevant findings being associated with regions as opposed to single sites [25].

It is likely for these reasons, methods designed to define or identify regions have proven popular, particularly on alternative array platforms. The "bump hunting" method of Jaffe et al [25], primarily intended for use with data from the NimbelGen CHARM 1M and 2M arrays [26], detects regions of interest, defined as those where methylation levels of a set of spatially contiguous probes associate with the outcome (assessed by linear modelling). There have also been methods developed specifically for 450K data: Ong and Holbrook's region detection method works similarly to bump hunting, defining candidate regions related to the outcome, this time starting with at least two consecutive probes, and then using a sliding windows approach to expand regions by adding further probes to the region [19]. The non-homogenous hidden markov model (NHMM) approach of Kuan and Chiang incorporates dependence of adjacent probes into a multiple testing framework [20].

Regional analysis has a couple of major limitations however which might explain why many researchers prefer site level analysis, and why reported associations still tend to be for individual CpGs. Regional methods require that the data possesses certain characteristics; for example, often, assumptions are made about the level of correlation between sites. As such, these methods are optimised for data from a specific platform, taking into account the distance between probes and evenness of coverage. For example, bump hunting when applied to 450K data only covers around 20% of the CpGs profiled on the array [19]. Even when the method is designed for the specific platform in question, there is no guarantee that all CpGs will be placed in regions. This is because CpG sites are distributed non-uniformly throughout the genome in regions of both high and low density, therefore, it is

expected that certain sites will not be located in any correlated block. Ong and Holbrook [19] found that their method, though designed for the 450K, does not include 24% of the probes on the array, leading the authors to suggest that single probe analysis should be performed along side region discovery to maximize discovery of differentially methylated sites.

4.1.3 The multiple testing problem

Whether an analysis aims to identify DMPs, DMRs, or both, because of the large number of simultaneous tests being performed, an over abundance of false positive results is expected at nominal thresholds. This problem is complicated by the presence of a correlation structure in the data, creating dependence between tests, and meaning that the actual number of independent tests carried is less than the total number of tests. There are several well-established strategies for controlling the Type I error rate in genetic association studies, which suffer from the very same issues. These are now discussed in turn.

4.1.4 FWER solutions to multiple testing

A well-established approach for dealing with multiple testing under independence is to control the family wise error rate (FWER), the probability of obtaining one or more significant results under the null of no association. Typically, this is achieved by approximating a per-test significance level α' corresponding to a FWER α for n number of tests, using the Bonferonni correction ($\alpha' = \frac{\alpha}{n}$), or the more exact Šidák correction ($\alpha' = 1 - (1 - \alpha)^{\frac{1}{n}}$) [27].

In the GWAS setting, control of the FWER by such methods is rarely carried out in practice, as the estimated per-test significance level can be overly conservative for rejecting individual hypotheses, and can increase the Type II error rate, reducing power [28]. One of the reasons that such methods fail to produce desirable results in GWAS is because the assumption of independence between tests is violated, due to the presence of LD across the genome. This very same issue requires careful con-

sideration in EWAS. The methylome has multiplicity of roughly the same order of magnitude as the genome with approximately 28 million CpG sites that could potentially be interrogated (although current technologies like the 450K measure far fewer). As discussed earlier, it has also been shown that there exists correlation between CpGs in close proximity. The presence of correlation between sites has led some to suggest that a significance threshold based on Bonferonni correction would similarly be overly conservative for EWAS [4, 29], and that FDR or permutation solutions might be more suitable [30].

4.1.5 FDR solutions to multiple testing

To overcome some of the problems associated with FWER correction, false discovery rate (FDR) methods can be employed. The FDR method as first proposed by Benjamini and Hochberg [31] (or see [32] and [33] for more recent extensions of the method) aims to control the proportion of Type I errors amongst the significant results (rejected null hypotheses). The method first sorts the p-values in ascending order, and then determines the number of tests (k) with a p-value of less than α ($FDR = \frac{n}{\alpha/k}$), giving the per-test significance level required to obtain an overall FDR of the level desired, for example 5%. FDR assumes independence between tests, but the method has been extended to account for complex dependence structures between tests [34].

Controlling the error rate using FDR methods can increase power, enabling a higher number of associations to be identified [35]. This increased power can be achieved by adjusting the error rate depending on the nature of the study (i.e. exploratory or confirmatory), the number of tests being performed, and expected number of true associations [36]. The intuitive definition of a FDR and the increased power it offers for detecting true associations, has made it popular for genetic studies. FDR is also finding application in EWAS, with all of the previously mentioned regional methods designed to identify DMRs [25, 20, 19] utilising FDR in their significance estimations.

There are a number of caveats to using FDR correction. Firstly, because FDR and

FWER are not equivalent, the p-value as a measure of significance is not appropriate and instead an analogous measure the q-value [37] is often used (the minimum FDR required for a test to reach significance). Secondly, the FDR can be higher than traditional Type I error rates [35].Thirdly, controling FDR under dependence can be overly conservative. Overall, these conspire to make FDR more difficult to use and the results arguably harder to interpret, complicating meta-analysis and cross study comparison [38]. This last point is particularly relevant in the context of EWAS, given the variety of different platforms used, and the importance of independent replication.

4.1.6 Resampling solutions to multiple testing

Resampling methods such as permutation testing are well suited to dealing with multiplicity and dependence in GWAS [39], and also have the advantage of producing exact p-values, potentially making such an approach attractive for EWAS. Permutation procedures typically involve randomly shuffling phenotypic labels to generate an empirical distribution of a test statistic under the null hypothesis of no association between the trait and markers being tested. Adjusted p-values can then be obtained by comparing observed test statistics with the distribution of maximum test statistics from each permutation [40]. Since the genotypes are fixed, the correlation structure remains intact, and the FWER is then correct under dependence [41].

Permutation approaches have a few disadvantages, which have limited their application in GWAS to some extent, and these would equally apply to EWAS. Firstly, full permutation testing can be inefficient and computationally intensive [42]. This inefficiency is largely due to the null distribution only being valid for the dataset it was generated for, so even analysing a subset of the data would require complete re-computation [41]. Another limitation of permutation testing is that the multiplicity of the entire genome is not accounted for [35]. It has previously been argued that the responsible use of a p-value threshold in GWAS should include all polymorphisms across the genome [35], not just those that were actually tested - an argument which could equally apply to the testing of differentially methylated

CpG sites in the case of EWAS.

4.1.7 Permutation correction

An attractive alternative to full permutation testing is to use the results from a permutation test to estimate an effective number of tests, which can then be used to calculate a FWER α. This is the approach taken by Dudbridge and Gusnanto [35] for the estimation of a genome-wide significance level. The approach uses a permutation scheme to generate a distribution of minimum p-values from which to calculate a threshold for the dataset (the 5% point). A subsampling method is then used to extrapolate the results to an array with saturated density, in order to derive a genome-wide threshold accounting for genome-wide multiplicity. This method has the advantages of permutation testing, in terms of allowing for both multiplicity and dependence, and also the advantages of FWER, in that it provides a frequentist p-value threshold which can be applied to other studies.

4.1.8 Estimating a significance threshold for EWAS

At present, there is no consensus on what an appropriate significance threshold for EWAS might be, although several authors have made recommendations. As mentioned previously, Tsai et al. [30] have suggested that a Bonferroni-adjusted threshold of $\alpha = 10^{-7}$ (accounting only for those CpGs tested on the 450K) would be overly conservative, and recommend using FDR or permutation methods for controlling the error rate until a consensus is established. Rakyan et al. [4] have proposed a liberal threshold of $\alpha = 10^{-6}$, based on a hypothetical set of 500K CpG probes with sufficient and uniform spacing between probes such that independence of individual measurements is assumed. Although the authors also acknowledge that due to correlation between neighbouring CpG sites, a specific calculation will be required, and a stringent level is more likely to lie between 10^{-8} and 10^{-7}.

The current research hopes to address the perceived need for an empirically derived

Table 4.1. Details about the five 450k datasets used.

Dataset	GEO accession	Population	Tissue	Age	Status	N
Gambian	GSE59592	African - Gambian	Blood	2-8 months	Healthy	120
CRC	N/A	Caucasian - European	Colon/rectum	58-80 years	Cancer	18
Caucasian	GSE40279	Caucasian - European	Blood	19-101 years	Healthy	426
Afr-Am	GSE41826	African - American	Brain	13-48 years	Healthy	12
Cau-Am	GSE41826	Caucasian - American	Brain	13-79 years	Healthy	65

p-value threshold for tests of differential methylation status of individual CpGs by applying the permutation testing based correction approach of Dudbridge and Gusnanto [35] mentioned above and described in further detail in the Methods section. This is applied to five distinct 450K datasets from a variety of study populations in order to derive specific significance thresholds. Further, in order to account for all CpG sites across the genome and not only those covered on the 450k, a subsampling procedure is used to extrapolate the results to a hypothetical array with infinite CpG site density and provide a more generally applicable genome-wide significance threshold.

4.2 Methods

4.2.1 Methylation datasets

The methylation data was derived from five independent studies utilizing the Illumina Infinium HumanMethylation450k platform, primarily chosen for their diversity in the populations studied and conditions under consideration, see Table 4.1. Two of these were recent methylomic profiling studies by collaborating groups who kindly allowed us to use the data in this study, and the remaining three were taken from publicly available datasets in the NCBIs Gene Expression Omnibus repository (http://www.ncbi. nlm.nih.gov/geo/), which were accessed and downloaded via the *marmal-aid* methylation database (http://marmal-aid.org) [43]. These public datasets were at the time of searching the largest three studies comprising healthy controls from Caucasian and African populations - which for the purpose of repli-

cation, were selected for their similarity to the populations studied in our own datasets.

For all datasets, the data was used in both $Beta$ value and M value form. To assess the correlation structure, $Beta$ values were used, as this is the form that 450K data is most commonly deposited into public databases. For permutation testing, M values were used, as these have previously been shown to provide better performance in terms of detection rate and true positive rate in differential methylation analysis [44]. In order to convert between $Beta$ and M values the following relationships were used (from [44]):

$$Beta_i = \frac{2^{M_i}}{2^{M_i} + 1} \tag{4.1}$$

$$M_i = log_2\left(\frac{Beta_i}{(1 - Beta_i)}\right) \tag{4.2}$$

4.2.1.1 Gambian

The first dataset, from here on referred to as the "Gambian" dataset, comes from an investigation into the effect of *in utero* exposure to aflatoxin B1 on embryonic development in a mother/child cohort from the Gambia [45], where peripheral blood samples from 120 infants between 3-6 months of age were used for methylation profiling. This data was provided in the form of a matrix of processed $Beta$ values, which had undergone QC and normalization. Full details on the experimental protocol and analysis method can be found in [45].

4.2.1.2 CRC

The second dataset, "CRC", is taken from a study characterising methylation patterns in 18 cases of colorectal carcinoma and 4 control samples of intestinal mucosa from a European caucasian population (unpublished work). An Illumina GenomeStudio report was provided which contained raw probe intensities, this was first processed using the *methylumi* package in *R* [46] in order to perform color

balance adjustment and quantile normalisation, and to generate a matrix of $Beta$ values. Control samples were removed due to the high level of discordance expected between cases and controls - a result of abberant CpG island methylation [47] and other large scale methylomic alterations expected in the CRC cases [48]. The resulting dataset contained only the 18 cases.

4.2.1.3 Public datasets: Caucasian/Afr-Am/Cau-Am

A further three datasets were identified by searching the *marmal-aid* database for healthy controls from Caucasian or African populations. This search identified two datasets. The first, "Caucasian", comes from a study into age related changes to methylomic state as profiled in peripheral blood samples from 426 caucasian individuals, spanning a wide age range [49] (GEO accession number: GSE40279). The second, yielding subsets "Afr-Am" and "Cau-Am", are from an unknown study (missing annotation in *marmal-aid*) consisting of 12 African-American and 65 Caucasian-American controls. For all of these *marmal-aid* datasets, processed $Beta$ values were used which had undergone quantile normalization and imputation of missing probes. This data was taken forward for analysis without any further processing.

4.2.2 Correlation structure

Prior to carrying out permutation testing, the correlation structure within each methylation dataset was assessed qualitatively. This was performed firstly to confirm previous findings demonstrating correlation between adjacent CpGs, and secondly, to investigate any potential differences in correlation structure between the datasets, which could for example be attributable to the different tissues, ethnicities, or diseases studied.

For each dataset, the following procedure was carried out to determine the level of correlation between adjacent probes. Using the subset of 46K probes mapping to chromosome 1, Pearson's correlation between each of the probe $Beta$ values was

calculated. Next, pairwise inter-probe distances were calculated by taking the absolute differences between the genomic positions of the CpGs (in bp) as given in the 450K annotation package in *R* [50]. This list of inter-pair distances was reduced by retaining only those with distance less than or equal to 10,000 bp. These remaining probe pairs were then binned into approximately 10,000 bins containing around 400 pairs each, and the median pairwise inter-probe distance of the bin recorded. The mean pairwise correlation of $Beta$ values per-bin was also recorded.

4.2.3 Permutation scheme

For each dataset, a permutation scheme was used to generate an empirical null distribution of t-test values, from which a per-CpG significance level α could be derived. Phenotypic labels were randomly assigned to samples, for sample size $n : \lfloor n/2 \rfloor$ were designated as cases and $\lceil n/2 \rceil$ as controls. The labels were then randomly permuted 10,000 times, and for each of these permutations, independent unrelated sample t-tests were performed for each CpG and the absolute t values recorded. Details of the algorithm are provided in Figure 4.1. Both this and the subsampling methods were implemented as command-line tools in *python* making use of multi-processing (via the standard multiprocessing library) and compressed, memory-mapped files (using the HDF5 - high density file format) in order to reduce computation time and memory requirements. The code is given in Appendix C - Figure C1, and these are also available to download from https://github.com/asaffa/PhD.

```
#Input : x = m x n matrix of methylation M values, where m = probes, n = subjects.
#Output : y = p x m matrix of absolute t values, p = permutations, m = probes.

#1. Phenotypes (case v control) randomly shuffled 10,000 times

for n in subjects:
Extract case status from string, store as boolean in vector v
for permutation p in 0 to 10,000, p = p + 1:
        u[p] <- shuffle(v)

#2. Perform independent t-test for all probes in each permutation,
#store in results matrix y

for permutation u 0 to 10,000, u = u + 1:
        for probe m in 0 to length(m), m = m + 1:
                tt <-  two-sample_ttest(x[m,u["cases"] == True],x[m,u["controls"] == True])
                y[p,m] <- absolute(tt)
```

Figure 4.1. Implementation details for permutation testing procedure

In order to estimate α corresponding to the significance threshold at the density of CpGs present on the 450K array, the maximum t-test statistic scores for each permutation were taken and corresponding p-values were calculated (under assumption of equal variance), and α taken as the 5% point of the distribution of minimum p-values.

4.2.4 *Effective number of tests*

The effective number of tests is the number of m independent tests required to obtain the observed α using the Bonferroni correction:

$$m = 0.05/\alpha \tag{4.3}$$

this was calculated for each α derived from the permutation data. To test whether the observed data was consistent with there being an effective number of independent tests, for each set of permutation min p-values, a beta distribution was fitted. If there is an effective number of tests then the minimum p-values from the permutation replicates should follow a beta distribution with parameters:

$$\beta(a = 1, b = m') \tag{4.4}$$

corresponding to the Šidák correction:

$$1 - (1 - \alpha)^{m'} \tag{4.5}$$

The parameters a and b of the β distribution can be estimated using the following method of moments estimators:

$$\hat{a} = \bar{x}(\frac{\bar{x}(1-\bar{x}}{s^2}) - 1) \tag{4.6}$$

$$\hat{b} = (1 - \bar{x})(\frac{\bar{x}(1-\bar{x})}{s^2} - 1) \tag{4.7}$$

fixing the value of the first parameter a to be 1, a starting estimate for the second parameter b can be obtained:

$$\hat{b} = \frac{1-\bar{x}}{\bar{x}} \tag{4.8}$$

Maximum log-likelihood estimation (MLE) of $\beta(1, \hat{b})$ was then performed using the *optim* function in *R* to obtain a final estimate of m' and this was compared to the value of m obtained from the Bonferonni equation given above.

4.2.5 Subsampling method

For each set of permutation results, to extrapolate the findings to an array of infinite density, a subsampling procedure was used. For each individual permutation in the permutation matrix, the p-values were sampled over a uniform grid of 100 densities (i.e. different fractions of CpGs) from 0 to 1, in increments of 0.01, and the minimum p-value at each density recorded. This procedure was repeated 100 times, and for each of these 100 replicates, the 5% point at each density across all the permutations was recorded. The mean 5% point for each density across all 100 replicates was then used in subsequent analysis. Further details of the algorithm are provided in Figure 4.2.

```
#Input : y = p x m matrix of p-values, p = permutations, m = probes.
#Output: z = p x o matrix of mean 5% values at each density for each permutation

#1. for each permutation, obtain minimum p-values across probe sampling densities
#(0.01 to 1), then calculate the 5% point for each density across all the permutations.
#Repeat this procedure 100 times to obtain a mean 5% value for each density across all
#of the permutations.

Initialise matrix rp to store results
for replicates r in 0 to 100, r = r + 1:
        Initialise matrix mp to store min p-values for each density across all permutations
        for each permutation p in 0 to 10,000, i = i + 1:
                Initialize vector vp to store min p-values for each density
                for each density (d in 0.01 to 1.00, d = d + 0.01):
                        shuffle(p)
                        sb <- sample(from = p, sample size = (d*length(p)) ,
                                                without replacement = True)
                        vp[d] <- min(of = sb)
                mp[p] <- vp
                 rp[r] <- apply(to = columns of mp, function = percentile(of = mp[cols],
                                        percentile point = 5), return = vector)
```

Figure 4.2. Implementation details for subsampling procedure

4.2.6 *Estimation of a genome-wide threshold*

At low densities, the CpGs are expected to be uncorrelated and hence independent. Increasing the density will hence increase the observed level of correlation between probes and according to the Bonferroni law, the 5% point should decrease in a manner inversely proportional to the density. Therefore, at high densities where the coverage is approaching saturation, the 5% point is expected to converge to an asymptote, which represents α for the entire genome.

To obtain estimates for asymptote, the adjusted p-value was calculated:

$$\alpha = 0.05/m \tag{4.9}$$

and the Monod function fitted. The Monod function was originally used to model the growth of microorganisms, but finds application in many different settings where growth is limited by resources (in this case, the growth in number of ef-

fective tests is limited by site density). The equation takes the form:

$$f(x;u,k) = \frac{ux}{(k+x)} \tag{4.10}$$

where u is the limit as $x \to \infty$, and k is the value for x for which

$$f(x) = u/2 \tag{4.11}$$

also known as the half saturation parameter. This function was fitted to the calculated values of m and the parameters estimated using least squares and genome-wide α estimated as:

$$\alpha = (0.05/u) \tag{4.12}$$

4.2.7 Sample size estimation

In a recent paper, Tsai and Bell [29] estimated sample size requirements for case-control and disease-discordant MZ twin design EWAS based on power simulations across a range of different sample and effect sizes. The simulations were performed using both t-test and Wilcoxon tests, for nominal (p = 0.05) and genome-wide ($p = 1 \times 10^{-6}$) significance levels. Because a new estimate of genome-wise α is derived here, we decided to re-calculate the sample sizes required for 80% power provided in Table 2. of the paper. To this end, the non-central t-distributions giving rise to the estimates were inferred, and then used to calculate sample sizes for the thresholds calculated above. The following relationship was used:

$$1 - \beta = 1 - F_{(n-1,ncp)}(t_{(1-\alpha/2)}) + F_{(n-1,ncp)}(-t_{(1-\alpha/2)}) \tag{4.13}$$

where $1 - \beta$ = 0.8, $F_{(n-1,ncp)}$ is the cumulative distribution function of the non-central t distribution with n-1 degrees of freedom and non-centrality parameter ncp, given by:

$$ncp = \sqrt{n} \times \frac{\delta}{\sigma} \tag{4.14}$$

and $t_{(1-\alpha/2)}$ the critical value when $\alpha = 1 \times 10^{-6}$. Treating as a minimum optimization problem (using the *optim* function in *R*),and solving σ across the range of sample sizes n and mean methylation differences δ yielded these estimates of σ. These estimates were then used in another optimization, this time solving to obtain sample sizes for each δ using the empirically derived estimates of α.

4.3 Results

4.3.1 Correlation structure

The correlation structure was assessed for each dataset, the line plots in Figure 4.3 show the overall pattern of correlation between methylation sites as a function of the distance between their genomic positions. The relationship between these variables appears consistent across datasets. The curves show a characteristic shape with pairs in close proximity having the highest average levels of correlation, which tails off sharply as distances approach 1KB, and then decreases more slowly from 1KB to 2KB, after which it appears to reach a limit just above zero, which is assumed to be the mean background level of correlation across the genome. In terms of the actual mean per-bin correlation values, these look comparable across the datasets, with the exception of the CRC set, with the approximate maximum per-bin correlations of 0.3, 0.55, 0.4, 0.35, 0.4 and approximate background levels of 0.07, 0.12, 0.04, 0.07, 0.07 for Gambian, CRC, Caucasian, Afr-Am, and Cau-Am sets respectively. There are also differences in the variability and spread of correlation values between neighbouring bins, perhaps indicating differences in apparent level of noise between sets.

4.3.2 Permutation and effective number of tests

The results from the permutation testing scheme are given in Table 4.2 . The 5% points of the minimum p distributions vary between the different datasets: Gambian $\alpha = 2.04 \times 10^{-7}$, CRC $\alpha = 3.53 \times 10^{-8}$, Caucasian $\alpha = 2.44 \times 10^{-7}$, Afr-

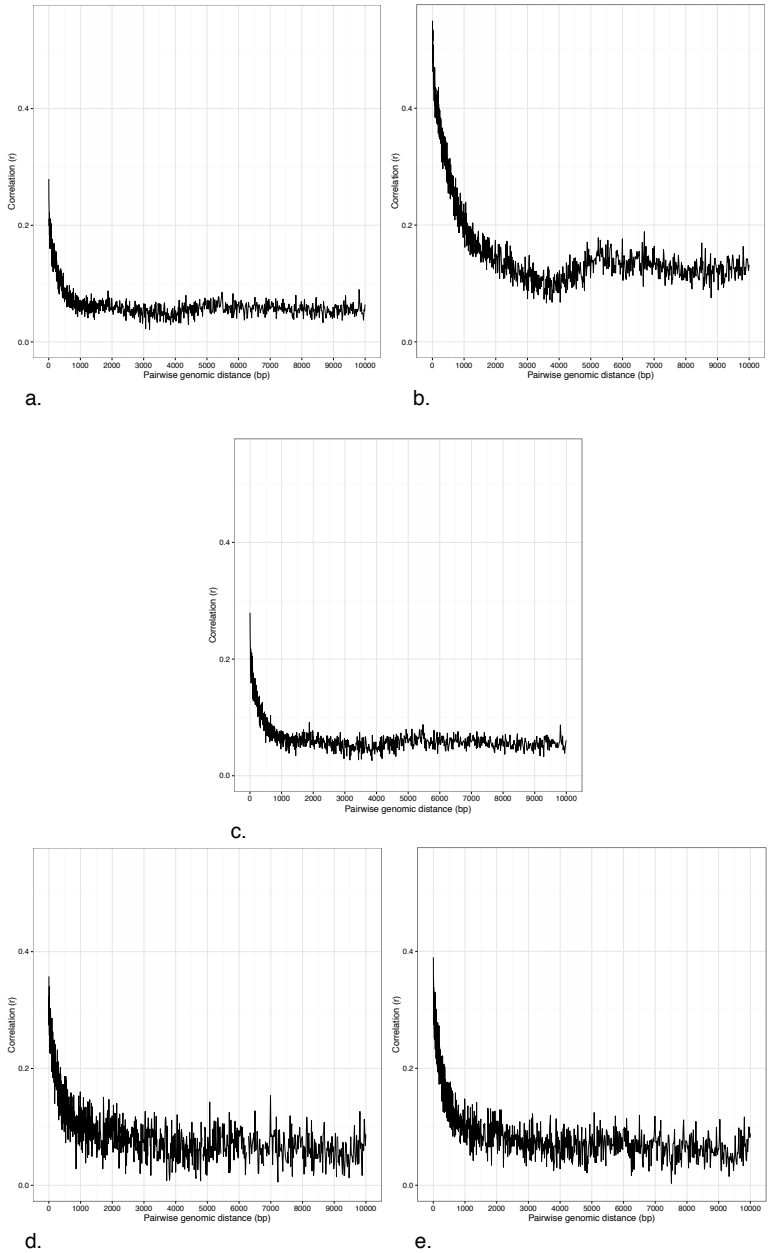

Figure 4.3. Correlation versus genomic distance for pairs of probes in chromosome 1. a. Gambian, b. CRC, c. Caucasian, d. Afr-Am, e. Cau-Am

Table 4.2. Permutation results showing the 5th percentile of the minimum p-values from 10,000 permutations of the datasets, m is the effective number of tests that this 5% point represents according to the Bonferroni law, and b is the estimated b parameter after fitting a $beta$ distribution.

Dataset	α	m	b
Gambian	2.04E-07	245563	170286
CRC	3.53E-08	1417410	161402
Caucasian	2.44E-07	204586	153670
Afr-Am	7.90E-08	633220	113038
Cau-Am	3.59E-07	139451	70782

Am $\alpha = 7.90 \times 10^{-8}$, Cau-Am $\alpha = 3.59 \times 10^{-7}$. The αs obtained for Gambia, Caucasian, and Cau-Am are larger than the Bonferonni adjusted 5% threshold of $\alpha = 1.07 \times 10^{-7}$, while the CRC and Afr-Am are smaller by almost a factor of 10. Investigating further, Figure 4.4 shows quantile-quantile plots of the observed minimum p distributions against the expected quantiles according to this Bonferonni threshold ($\beta(1, 467264)$). From these plots it can be observed for the Gambia, Caucasian, and Cau-Am sets, that while Bonferonni produces uniformly distributed p-values, these are also deflated over the entire range. For the CRC and Afr-Am sets, it appears that the observed data is not at all well modelled by the β distribution, and there is not a clear pattern of over or under correction of p-values, as they appear both deflated at low values and inflated at higher values.

Converting these 5% points to effective number of tests using Formula 1.9 gives: Gambian m = 245,563, CRC m =1,417,410, Caucasian m = 204,586, Afr-Am m = 633,220, Cau-Am m = 139,451. Fitting the β distribution (Formulæ1.4 to 1.8) then yields the following estimates for the b parameter: Gambian b = 170,286, CRC b = 161,402, Caucasian b = 153,670, Afr-Am b = 113,038, Cau-Am b = 70,782. These different estimates of number of effective tests are not in close agreement, suggesting that the β distribution with parameters $(1, m)$ does not adequately model the minimum p-value distribution for m tests. The results for CRC and Afr-Am in particular do not seem to produce the expected distribution of min p-values, and the estimated values for α, m and b show little resemblance to the values obtained for the other datasets, perhaps indicating issues with these two particular sets, for instance, both having small sample sizes.

Table 4.3. Results from subsampling, showing the final values for the u and k parameters after fitting the Monod function, and the asymptote representing the genome-wide significance threshold alpha.

Dataset	u	k	Genome-wide α
Gambian	1.38E+06	4.71E+00	3.61E-08
CRC	N/A	N/A	N/A
Caucasian	1.18E+06	4.79E+00	4.25E-08
Afr-Am	N/A	N/A	N/A
Cau-Am	6.38E+05	3.54E+00	7.83E-08

Assuming the results from correlation analysis are reliable and have been correctly interpreted, there are negligible differences between the overall patterns of co-methylation in the data from the different populations. Therefore, the estimated thresholds for the three groups producing permutation data most closely fitting the expected distribution, Gambian, Caucasian and Cau-Am, can be combined into a single figure. Taking the weighted mean of these different estimates yields a 450K-specific $\alpha = 2.5 \times 10^{-7}$.

4.3.3 Subsampling and genome-wide threshold

For subsampling, Figure 4.5 shows the resulting plots of the mean 5% of the sampled minimum p-values at each subsampling density. The asymptotes were estimated by fitting the monod function to the effective number of tests, the closeness of the fit can be seen in Figure 4.6. The resulting estimates for u, the limit as the density approaches infinity, and k, the value for x at which half of genome-wide multiplicity is accounted for, are given in Table 4.3. The asymptotes are shown in Figure 4.5. It was not possible to fit the monod function to the CRC and Afr-Am.

The estimates for u are: Gambian u = 1,380,000, Caucasian u = 1,180,000, and Cau-Am u = 638,000, which are used to calculate genome-wide corrected significance thresholds, giving: Gambian $\alpha = 3.61 \times 10^{-8}$, Caucasian $\alpha = 4.25 \times 10^{-8}$, and Cau-Am $\alpha = 7.83 \times 10^{-8}$. Once again, assuming the results from correlation assessment are correct and that there are minimal differences in the correlation structure be-

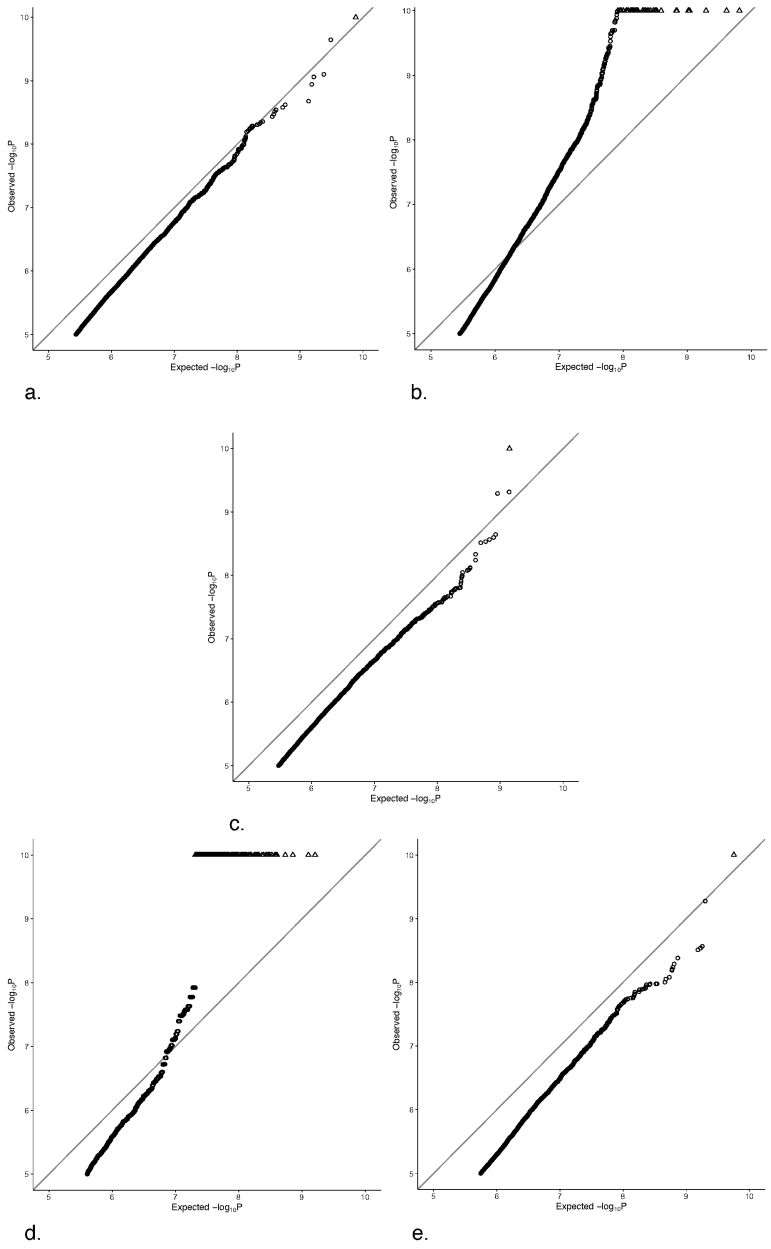

Figure 4.4. QQ plots showing observed distribution of minimum p values verses the expected distribution under complete independence. a. Gambian, b. CRC, c. Caucasian, d. Afr-Am, e. Cau-Am

tween the different populations, these can be combined to give a weighted mean of genome-wide $\alpha = 4.5 \times 10^{-8}$.

4.3.4 Sample size estimation

	Twin				Case-control			
Diff	$p < 0.05$	$p < 1$ x 10-6	$p < 2.5$ x 10-7	$p < 4.5$ x 10-8	$p < 0.05$	$p < 1$ x 10-6	$p < 2.5$ x 10-7	$p < 4.5$ x 10-8
7	30	178	195	216	37	211	231	256
8	25	145	159	176	30	169	185	205
9	20	117	128	142	24	137	150	166
10	17	98	107	119	20	112	123	136
11	15	81	89	99	17	96	105	117
12	13	71	78	86	15	80	88	97
13	11	63	69	77	13	70	77	85
14	10	55	60	67	11	61	67	74
15	9	50	55	61	10	54	59	66

Table 4.4. Sample size estimates based on those presented in [29] using the estimates for 450k and genome-wide significance derived in this study. Diff is the percentage mean methylation difference between case and control, twin and case-control refer to the study designs, and for each of the significance thresholds,the sample sizes required to achieve a power of 0.8 are given.

Results for sample size calculations are given in 4.4 alongside the original results from [29]. Using the 450K-specific $\alpha = 2.5 \times 10^{-7}$, increases the sample size estimates by ~10% over those estimated from power simulations using the suggested threshold of $\alpha = 1 \times 10^{-6}$. With the estimated genome-wide $\alpha = 4.5 \times 10^{-8}$, the sample size estimates are increased by ~20%.

4.4 Discussion

4.4.1 Correlation structure

In order to determine whether any differences in the overall correlation structure between the different datasets exists, we investigated the relationship between co-methylation and genomic distance between pairs. The results reveal a distinctive relationship, whereby proximal sites, up to a distance of around 1K bases apart, show a moderate level of correlation in the 0.25 to 0.4 range, falling to background levels once inter-pair distances reach around 2KB. These results are consistent with

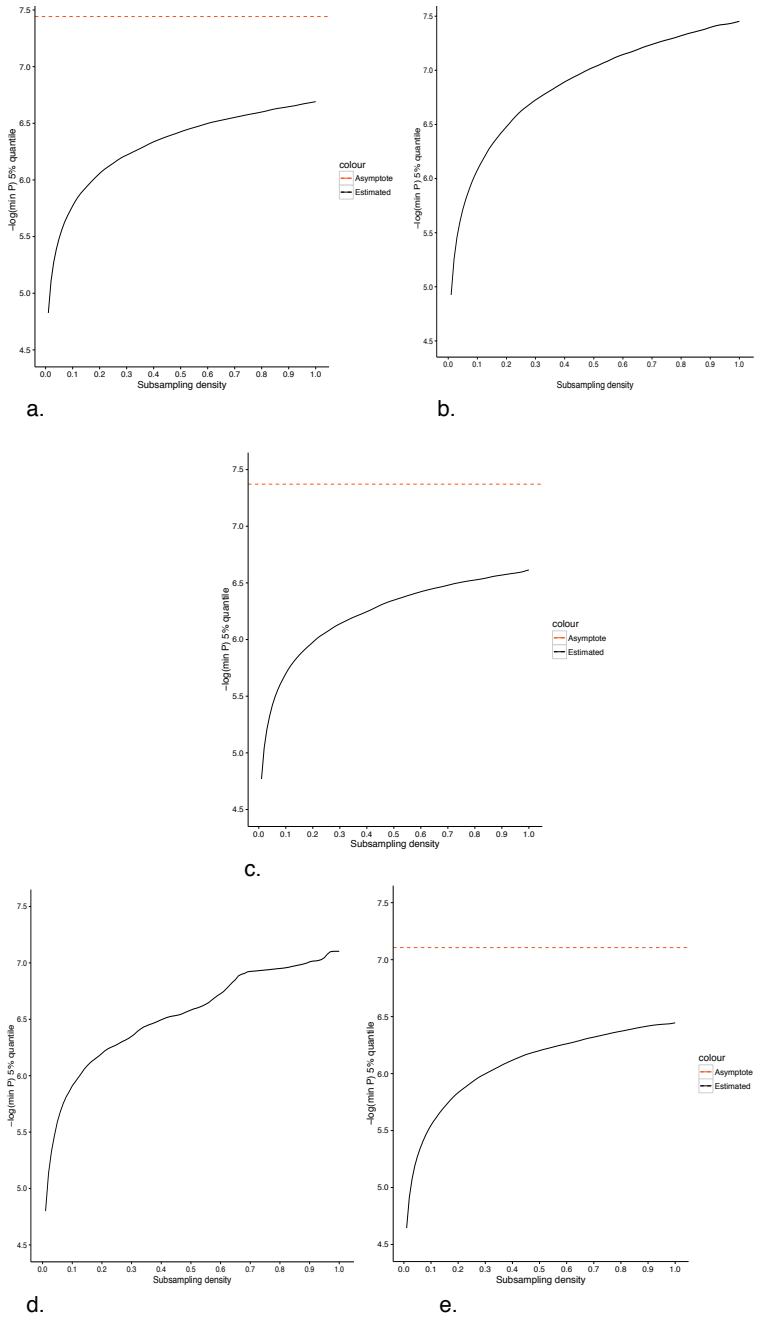

Figure 4.5. Significance threshold as a function of CpG site density following the subsampling procedure. Where possible, the monod function was fitted to estimate an asymptote representing the threshold at fully saturated CpG density, this is indicated by a dashed red line. a. Gambian, b. CRC, c. Caucasian, d. Afr-Am, e. Cau-Am.

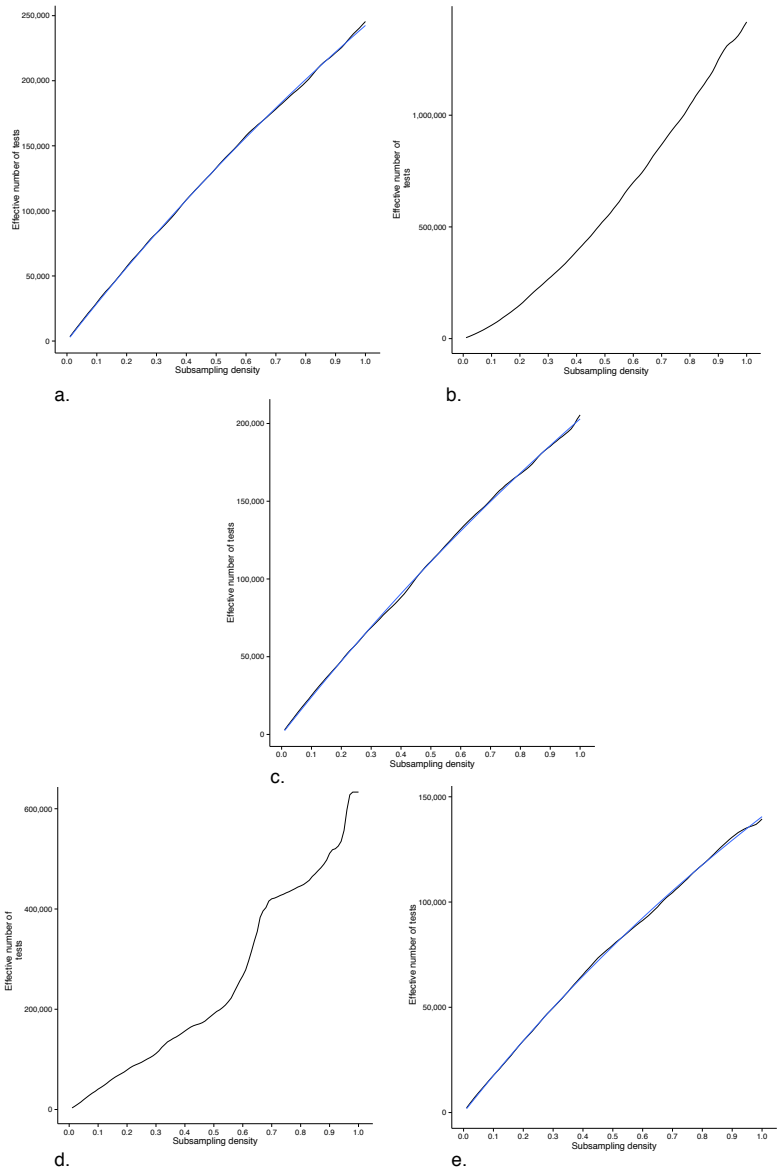

Figure 4.6. Estimated number of tests as a function of CpG site density. Where possible, the monod function was fitted, this fit is shown by the blue line. a. Gambian, b. CRC, c. Caucasian, d. Afr-Am, e. Cau-Am.

previous findings, which have found moderate to strong correlations of between 0.26 [25] and 0.45 [19] extending over genomic distances of between 1 and 2 KB [18, 25, 20, 19].

There do not appear to be any large-scale differences in overall patterns of co-methylation between brain and blood when comparing the Cau-Am and Caucasian datasets, or between Gambian and Caucasian populations when comparing the Gambian and Caucasian datasets. Only the CRC dataset shows some slight differences. While it would be tempting to link this to the disease, this could equally be due to data heterogeneity, especially given that all the datasets are taken from separate studies, conducted by different labs, using different experimental protocols. Differences in the variability between sets are also apparent, with the spread of correlation values between adjacent bins of pairs showing a greater spread for the datasets having smaller sample sizes. Once again, this is merely speculated, as it is not possible to rule out the differences being purely the result of data hetereogeneity.

As far as we are aware, this is one of the first attempts to characterise the overall patterns of co-methylation between CpGs in the context of different tissues, ethnicities and disease states. The initial indications here are there are unlikely to be any large-scale differences in terms of the overall relationship between adjacent probes across tissues and ethnicities considered. It is possible that there are differences in the case of certain cancers such as CRC. Further experimental work would be required before any more general conclusions could be made, preferably involving the generation of new data under rigorously controlled conditions in order to minimise confounders such as sample selection (especially controlling for age and smoking), sample collection procedures, sample composition, and batch effects, and to reduce variablilty in data processing.

4.4.2 450K threshold

We used a permutation scheme to obtain values for α for each dataset. Inspecting the fit of the β distribution $\beta(1, 467264)$ to the minimum p-values generated, the

datasets CRC and Afr-Am deviate from the expected distribution quite substantially, perhaps a sign that these datasets had not produced reliable results, possibly due to their small sample sizes. For the Gambia, Caucasian, and Cau-Am, the β function seems to produce inflated p-values consistently inflated over the range, which suggests that reasonable fit might be acheived by adjusting the second parameter of the β distribution, in other words, choosing a different value for the effective number of tests in the Bonferonni equation.

As for the actual 5% values, for the Gambia, Caucasian, and Cau-Am sets these are all larger than the Bonferonni adjusted 5% threshold of $\alpha = 1.07 \times 10^{-7}$. Assuming that the results and conclusions from the assessment of correlation within the different sets are correct, and that there are no major differences between different populations, then the results from these three sets can be combined to give a weighted mean of $\alpha = 2.5 \times 10^{-7}$. From permutation testing results, we conclude that a significance threshold of $\alpha = 2.5 \times 10^{-7}$ would be appropriate for the 450K, accounting for the subset of probes tested on the array but not the hypothetical set of probes that could be tested with fully saturated genome-wide coverage.

4.4.3 Genome-wide threshold

To address the issue of genome-wide multiplicity, we used a subsampling method to extrapolate the results of permutation testing to an array of infinite density. The results of fitting the Monod function to the subsampling data revealed that the limit for the number of tests as the coverage on the array becomes saturated is somewhere in the region of 1×10^6, i.e. 1M probes, which is 4 times the current density. The results from the three different sets were combined, giving genome-wide $\alpha = 4.5 \times 10^{-8}$. Interestingly, this figure is similar to that typically used for GWAS : $\alpha = 5 \times 10^{-8}$, although it is not clear why this would be the case considering they are different molecular modalities with their own particular characteristics. Further investigation would be required to determine whether this is merely coincidence.

Comparing this with previous recommendations, we see that the estimate for a

genome-wide significance for EWAS obtained here is smaller than what would be considered a liberal threshold of 10^{-6} [4], and would fall within the range considered stringent 10^{-8} to 10^{-7} [4]. This threshold is however much less stringent than the genome-wide Bonferonni level, which assuming there are 28 million CpGs across the genome, is $\alpha = 1.79 \times 10^{-9}$. Bonferonni would then indeed be overly conservative for methylation data, but perhaps not to the extent previously suggested.

As for the limitations of this method, permutation correction attempts to identify a limit for number of independent tests using a very small sample. The 450K offers only ~2% coverage of the methylome, from which we have attempted to extrapolate a more general relationship between the 5% point of the minimum p-values and the density of coverage. Further complicating matters is the design of the 450K array itself, which has been described as offering relatively sparse and irregular coverage [19], perhaps making it a less than ideal data type for such an approach.

Related to issues around extrapolation, is the question of whether any estimated genome-wide threshold would be more generally applicable in EWAS, for DNA methylation measurements generated from different tissues, different populations, using different array platforms, or even different technologies such as MeDIP-seq (methylated DNA immunoprecipitation sequencing), WGBS (whole genome bisulfite-sequencing), or RRBS (reduced representation bisulfite sequencing). In terms of cross-tissue applicability, given the now wide-spread availability of public 450k datasets, further work could be done in comparing thresholds derived for different tissues. In a similar vein, we could also repeat the study for different populations, or perhaps even perform a huge meta-analysis using all healthy controls from every available study (which it appears might be possible to do in the latest version of *marmalaid*). As for cross-platform applicability, we would expect that a genome-wide threshold could be used irrespective of the platform or level of coverage, with the caveat that any estimation should ideally be made based on data from a platform with even, un-biased coverage of all features to increase confidence in the extrapolation. Such a threshold could even find application in whole-genome approaches such as WGBS where even though complete coverage of the methylome can be achieved, because permutation testing is computationally expensive,

time consuming, and study specific, an a-priori estimate (even though it will be less accurate) could still be desirable.

4.4.4 Sample size estimation

A previous study by Tsai and Bell [29] used power simulations to estimate sample sizes required to detect a range of mean methylation differences for both twin and case control designs based on nominal significance of $\alpha = 0.05$ and an estimated genome-wide threshold of $\alpha = 1 \times 10^{-6}$. By inferring the parameters of the t-distributions giving rise to these estimates, we re-calculated sample sizes for our empirically derived 450K and genome-wide thresholds. The results indicated that using the 450K-specific $\alpha = 2.5 \times 10^{-7}$ would require sample sizes ~10% larger than those previously estimated, and using the genome-wide $\alpha = 4.5 \times 10^{-8}$ would require samples ~20% larger, in order to obtain the same power.

4.4.5 Conclusion

There are many reasons to prefer FDR or permutation methods over Bonferonni type control of the FWER - especially if the identification of DMRs is the analytical goal, however, there is still value in obtaining a genome-wide significance level for individual DMPs. Firstly, a significant global test can be taken as evidence that there is a sufficiently strong signal present in the data to differentiate true from false positive [35], which can inform downstream analytical decisions, including whether to proceed with a site-specific or regional analysis. Secondly, because EWAS are often conducted as exploratory studies, it is useful to have a method for ranking associations to base decisions on which should be experimentally replicated; the FWER adjusted significance provides a simple, statistically sound method to do so. Thirdly, the use of a standard threshold also aids comparison across experiments, simplifying meta-analysis. Finally, having an *a priori* threshold estimate enables power calculations to be made, which can aid in the design of future experiments.

The $\alpha = 2.5 \times 10^{-7}$ threshold for significance for the 450K array that we have estimated here takes into account both dependency between test due to patterns of co-methylation, and the multiplicity of the set of CpGs tested. As it is derived from results averaged over European, African, and American populations, it could therefore find general application for 450K methylation data, offering an empirically derived, more permissive alternative to FWER correction. One major limitation of the 450K threshold is that it does not take into account genome-wide multiplicity, although the impact of this on the accuracy of the estimate has not been possible to determine. In addition, the non-random placement of probes and the current limited coverage of the 450K makes extrapolation to saturated probe coverage somewhat less reliable, reducing confidence in the derived genome-wide $\alpha = 4.5 \times 10^{-8}$.

With the recent release of Illumina's 850K EPIC array, with almost double the coverage of the 450k, the issue of significance in EWAS is as relevant as ever. Future work may then seek to apply the methods outlined here to similarly derive significance threshold estimates for the EPIC array, and compare with the findings of this study. It is hoped doing so may go some way to addressing the question of whether it will be possible to establish a universal, platform-agnostic EWAS threshold for single site-level differential methylation analysis.

References

[1] A. Bird, "Dna methylation patterns and epigenetic memory," *Genes & development*, vol. 16, no. 1, pp. 6–21, 2002.

[2] K. D. Robertson, "Dna methylation and human disease," *Nature Reviews Genetics*, vol. 6, no. 8, pp. 597–610, 2005.

[3] V. K. Rakyan and S. Beck, "Epigenetic variation and inheritance in mammals," *Current opinion in genetics & development*, vol. 16, no. 6, pp. 573–577, 2006.

[4] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck, "Epigenome-wide association studies for common human diseases," *Nature Reviews Genetics*, vol. 12, no. 8, pp. 529–541, 2011.

[5] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450k technology," *Epigenomics*, vol. 3, no. 6, pp. 771–784, 2011.

[6] H. Heyn, F. J. Carmona, A. Gomez, H. J. Ferreira, J. T. Bell, S. Sayols, K. Ward, O. A. Stefansson, S. Moran, J. Sandoval, *et al.*, "Dna methylation profiling in breast cancer discordant identical twins identifies dok7 as novel epigenetic biomarker," *Carcinogenesis*, vol. 34, no. 1, pp. 102–108, 2013.

[7] K. Walter, T. Holcomb, T. Januario, P. Du, M. Evangelista, N. Kartha, L. Iniguez, R. Soriano, L. Huw, H. Stern, *et al.*, "Dna methylation profiling defines clinically relevant biological subsets of non–small cell lung cancer," *Clinical Cancer Research*, vol. 18, no. 8, pp. 2360–2373, 2012.

[8] Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, *et al.*, "Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis," *Nature biotechnology*, vol. 31, no. 2, pp. 142–147, 2013.

[9] E. Swan, A. Maxwell, and A. McKnight, "Distinct methylation patterns in genes that affect mitochondrial function are associated with kidney disease in blood-derived dna from individuals with type 1 diabetes," *Diabetic Medicine*, 2015.

[10] J. I. Feinberg, K. M. Bakulski, A. E. Jaffe, R. Tryggvadottir, S. C. Brown, L. R. Goldman, L. A. Croen, I. Hertz-Picciotto, C. J. Newschaffer, M. D. Fallin, *et al.*, "Paternal sperm dna methylation associated with early signs of autism risk in an autism-enriched cohort," *International journal of epidemiology*, p. dyv028, 2015.

[11] Y. Song, K. Miyaki, T. Suzuki, Y. Sasaki, A. Tsutsumi, N. Kawakami, A. Shimazu, M. Takahashi, A. Inoue, C. Kan, *et al.*, "Altered dna methylation status of human brain derived neurotrophis factor gene could be useful as biomarker of depression," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 165, no. 4, pp. 357–364, 2014.

[12] E. Walton, J. Hass, J. Liu, J. L. Roffman, F. Bernardoni, V. Roessner, M. Kirsch, G. Schackert, V. Calhoun, and S. Ehrlich, "Correspondence of dna methylation between blood and brain tissue and its application to schizophrenia research," *Schizophrenia bulletin*, p. sbv074, 2015.

[13] I. Florath, K. Butterbach, H. Müller, M. Bewerunge-Hudler, and H. Brenner, "Cross-sectional and longitudinal changes in dna methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated cpg sites," *Human molecular genetics*, vol. 23, no. 5, pp. 1186–1201, 2014.

[14] B. R. Joubert, S. E. Håberg, D. A. Bell, R. M. Nilsen, S. E. Vollset, Ø. Midttun, P. M. Ueland, M. C. Wu, W. Nystad, S. D. Peddada, *et al.*, "Maternal smoking and dna methylation in newborns: in utero effect or epigenetic inheritance?,"

Cancer Epidemiology Biomarkers & Prevention, vol. 23, no. 6, pp. 1007–1017, 2014.

[15] B. I. Laufer, J. Kapalanga, C. A. Castellani, E. J. Diehl, L. Yan, and S. M. Singh, "Associative dna methylation changes in children with prenatal alcohol exposure," *Epigenomics*, no. 0, pp. 1–16, 2015.

[16] K. B. Michels, A. M. Binder, S. Dedeurwaerder, C. B. Epstein, J. M. Greally, I. Gut, E. A. Houseman, B. Izzi, K. T. Kelsey, A. Meissner, *et al.*, "Recommendations for the design and analysis of epigenome-wide association studies," *Nature methods*, vol. 10, no. 10, pp. 949–955, 2013.

[17] J. Mill and B. T. Heijmans, "From promises to practical strategies in epigenetic epidemiology," *Nature Reviews Genetics*, vol. 14, no. 8, pp. 585–594, 2013.

[18] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, *et al.*, "Dna methylation profiling of human chromosomes 6, 20 and 22," *Nature genetics*, vol. 38, no. 12, pp. 1378–1385, 2006.

[19] M.-L. Ong and J. D. Holbrook, "Novel region discovery method for infinium 450k dna methylation data reveals changes associated with aging in muscle and neuronal pathways," *Aging Cell*, vol. 13, no. 1, pp. 142–155, 2014.

[20] P. F. Kuan and D. Y. Chiang, "Integrating prior knowledge in multiple testing under dependence with applications to detecting differential dna methylation," *Biometrics*, vol. 68, no. 3, pp. 774–783, 2012.

[21] E. L. Meaburn, L. C. Schalkwyk, and J. Mill, "Allele-specific methylation in the human genome: implications for genetic studies of complex disease," *Epigenetics*, vol. 5, no. 7, pp. 578–582, 2010.

[22] Y. Liu, X. Li, M. J. Aryee, T. J. Ekström, L. Padyukov, L. Klareskog, A. Vandiver, A. Z. Moore, T. Tanaka, L. Ferrucci, *et al.*, "Gemes, clusters of dna methylation under genetic control, can inform genetic and epigenetic analysis of disease," *The American Journal of Human Genetics*, vol. 94, no. 4, pp. 485–495, 2014.

[23] K. Schildknecht, S. Olek, and T. Dickhaus, "Simultaneous statistical inference for epigenetic data," *PloS one*, vol. 10, no. 5, p. e0125587, 2015.

[24] M. D. Robinson, A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber, and X. Zhou, "Statistical methods for detecting differentially methylated loci and regions," *Frontiers in genetics*, vol. 5, 2014.

[25] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International journal of epidemiology*, vol. 41, no. 1, pp. 200–209, 2012.

[26] R. A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddeloh, B. Wen, and A. P. Feinberg, "Comprehensive high-throughput arrays for relative methylation (charm)," *Genome research*, vol. 18, no. 5, pp. 780–790, 2008.

[27] Z. Šidák, "Rectangular confidence regions for the means of multivariate normal distributions," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 626–633, 1967.

[28] D. R. Nyholt, "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other," *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 765–769, 2004.

[29] P.-C. Tsai and J. T. Bell, "Power and sample size estimation for epigenome-wide association scans to detect differential dna methylation," *International journal of epidemiology*, p. dyv041, 2015.

[30] P.-C. Tsai, T. D. Spector, and J. T. Bell, "Using epigenome-wide association scans of dna methylation in age-related complex human traits," *Epigenomics*, vol. 4, no. 5, pp. 511–526, 2012.

[31] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[32] J. D. Storey, "The positive false discovery rate: a bayesian interpretation and the q-value," *Annals of statistics*, pp. 2013–2035, 2003.

[33] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, pp. 491–507, 2006.

[34] J. D. Storey and R. Tibshirani, "Estimating false discovery rates under dependence, with applications to dna microarrays," *Technical of Report 2001*, vol. 28, 2001.

[35] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genetic epidemiology*, vol. 32, no. 3, p. 227, 2008.

[36] K. J. Verhoeven, K. L. Simonsen, and L. M. McIntyre, "Implementing false discovery rate control: increasing your power," *Oikos*, vol. 108, no. 3, pp. 643–647, 2005.

[37] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 479–498, 2002.

[38] R. Higdon, G. van Belle, and E. Kolker, "A note on the false discovery rate and inconsistent comparisons between experiments," *Bioinformatics*, vol. 24, no. 10, pp. 1225–1228, 2008.

[39] A. P. Morris and L. R. Cardon, "Chapter 37 - whole genome association," in *Handbook of statistical genetics* (D. J. Balding, M. Bishop, and C. Cannings, eds.), vol. 1, ch. 37, pp. 1238–1263, John Wiley & Sons, 2008.

[40] G. A. Churchill and R. W. Doerge, "Empirical threshold values for quantitative trait mapping.," *Genetics*, vol. 138, no. 3, pp. 963–971, 1994.

[41] C. J. Hoggart, T. G. Clark, M. De Iorio, J. C. Whittaker, and D. J. Balding, "Genome-wide significance for dense snp and resequencing data," *Genetic epidemiology*, vol. 32, no. 2, pp. 179–185, 2008.

[42] F. Dudbridge and B. P. Koeleman, "Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies," *The American Journal of Human Genetics*, vol. 75, no. 3, pp. 424–435, 2004.

[43] R. Lowe and V. K. Rakyan, "Marmal-aid–a database for infinium human-methylation450," *BMC bioinformatics*, vol. 14, no. 1, p. 359, 2013.

[44] P. Du, X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis," *BMC bioinformatics*, vol. 11, no. 1, p. 587, 2010.

[45] H. Hernandez-Vargas, J. Castelino, M. J. Silver, P. Dominguez-Salas, M.-P. Cros, G. Durand, F. Le Calvez-Kelm, A. M. Prentice, C. P. Wild, S. E. Moore, *et al.*, "Exposure to aflatoxin b1 in utero is associated with dna methylation in white blood cells of infants in the gambia," *International journal of epidemiology*, p. dyv027, 2015.

[46] S. Davis, P. Du, S. Bilke, T. Triche Jr, and M. Bootwalla, "methylumi: Handle illumina methylation data," *R package version*, vol. 2, no. 0, 2012.

[47] J. F. Costello, M. C. Frühwald, D. J. Smiraglia, L. J. Rush, G. P. Robertson, X. Gao, F. A. Wright, J. D. Feramisco, P. Peltomäki, J. C. Lang, *et al.*, "Aberrant cpg-island methylation has non-random and tumour-type–specific patterns," *Nature genetics*, vol. 24, no. 2, pp. 132–138, 2000.

[48] K. Uhlmann, K. Rohde, C. Zeller, J. Szymas, S. Vogel, K. Marczinek, G. Thiel, P. Nürnberg, and P. W. Laird, "Distinct methylation profiles of glioma subtypes," *International Journal of Cancer*, vol. 106, no. 1, pp. 52–59, 2003.

[49] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, *et al.*, "Genome-wide methylation profiles reveal quantitative views of human aging rates," *Molecular cell*, vol. 49, no. 2, pp. 359–367, 2013.

[50] T. Triche Jr, "Illuminahumanmethylation450k. db: Illumina human methylation 450k annotation data," *R package version*, vol. 1, no. 6, 2012.

5 Integrating multi-dimensional omics datasets

5.1 Introduction

The rapidly falling cost of high-throughput sequencing technologies has led to a deluge in high-resolution omics data. It is now possible to generate data on DNA sequence variation, copy number variants, gene expression, non-coding RNA expression, DNA methylation levels and protein abundance, to name but a few. The increased computational requirements for storing, processing and accessing this data has been a major driving force behind the pace of innovation seen in bioinformatics. While some of the challenges associated with "big" data are likely to remain with us for the foreseeable future, recently attention has started to shift to another problem - how best to utilise this wealth of diverse molecular data to ask new questions, make new insights and inform our biological understanding. Integrative genomics approaches are being developed in order to tackle this problem.

Integrative genomics is primarily concerned with the development and application of methodologies for combining data from multiple omics datasets. The main motivation for doing so is firstly to validate findings from individual studies, and secondly, to gain a more holistic, higher-resolution, systems-level view of the genome and regulatory mechanisms involved in complex diseases. Integrative methods have been successfully applied to a multitude of problems, from survival analysis

in ovarian cancer [1], tumour subtype classification in glioblastoma [2], identification of key molecular drivers in chronic obstructive pulmonary disease [3], and discovery of disease susceptibility networks in type II diabetes [4].

5.1.1 An overview of integrative genomics

Perhaps due to the complex and interdisciplinary nature of the problem, the problem of data integration is not well defined and the literature is currently a little disparate [5]. Nonetheless, a number of reviews have attempted to provide a general overview of the area. In one of the earliest reviews, Hamid et al. [5] proposed a conceptual framework for data integration, which considers integration of both homogenous and heterogeneous data. The main biological questions that integrative approaches might be used to address were identified as: 1) differential expression, 2) copy number variation, 3) feature extraction, 4) disease classification, 5) gene mapping. Integration approaches were classified as being either 1) early, 2) intermediate, or 3) late referring to the level at which data is integrated, whether raw experimental data, transformed data, or summary-level data respectively. Some relevant statistical methods were discussed in detail, namely those relevant to horizontal integration (e.g. p-value combination).

Hawkins et al. [6] considered integration from the view of the "bench" scientist, discussing in terms of the types of experimental data available and the biological questions that might be addressed. The primary questions that can be addressed by integration were identified as, 1) annotation of functional features of the genome, 2) inferring function of genetic variants, and 3) gaining a mechanistic understanding of gene regulation. A brief overview of statistical and computational methods was given, being categorised as those concerned with data dimensionality reduction and supervised and unsupervised approaches, for classification type problems and inference intended to uncover relationships between datasets.

In a Nature Genetics review published in 2015, Ritchie et al. [7] examined data integration in an epidemiological context, where the analytical aim is to combine data in order to improve statistical models of complex traits. Two main approaches

are described: multi-staged analysis, which combines data two sets at a time in a stepwise fashion, and meta-dimensional analysis, where the different datasets are combined simultaneously and used build a multi-variate model that relates the molecular traits to the phenotypic trait of interest. Multi-staged analysis is further divided into genomic variation analysis, where genetic data is the starting point and the other datasets are used to functionally annotate SNPs (e.g. as eQTLs, mQTLs), and domain-knowledge guided analysis, where functional annotation is obtained from public repositories. Meta-dimensional analysis is divided into concatenation-based, transformation-based, and mode-based, depending on whether raw data is combined in the model, or if it transformed first before being modelled, or if datasets are modelled individually and then integrated. A table listing available bioinformatics software was also provided for each of these different approaches.

Kannan et al. [8] discussed data acquisition as opposed to utilization, describing some of the currently available large public data resources such as ENCODE (The Encyclopaedia of DNA Elements) [9] and FANTOM5 (Functional Annotation of Mammalian Genomes 5) [10]. There is a brief discussion of integration methods where three main categories are identified: 1. exploratory data analysis to identify intra and inter dataset patterns, 2. network analysis, to refine interaction networks by using molecular data, 3. supervised meta-analysis, where findings from one assay are refined by including further omics data. For each category, relevant bioinformatics tools were listed.

Finally, the textbook "Integrating omics data" by Tseng et al. [11] covers a wide range of topics in data integration in great detail, especially in relation to cancer studies. The distinction is made between two main classes of integrative problem, horizontal integration, which is the 'classic' meta-analysis of one molecular data type across multiple studies, and vertical integration, where multiple heterogeneous datasets on the same sample are combined, and. There are chapters outlining methods for discovering functional relevant modules, penalized regression methods for conducting joint analysis, Bayesian graphical models for integration, among others.

Taken together, these publications provide a comprehensive summary of the theoretical background, statistical methods or algorithms for particular classes of approaches, and available bioinformatics software to-date. However, we were not able to find any single review that adequately covered all three aspects from the perspective of neuropsychiatric or neurodevelopmental genetics, where there is considerable interest in combining heterogeneous datasets to help uncover the molecular etiology and pathomechanisms underlying conditions such as ASD.

5.1.2 Vertical integration

Our group are interested in vertical integrative strategies as means to dissect the molecular aetiology of cognitive and behavioural traits and neurodevelopmental disorders, using multi-assay measurements on the same individuals. There is a certain degree of overlap between the methods used in both horizontal and vertical integration, and a real world problem might include both strategies ("diagonal" integration), nonetheless, in this chapter we focus on vertical type integration. The rationale behind combining heterogeneous data is to leverage the available evidence in the most effective way, increasing precision, accuracy and power for identifying true signals [5]. Utilising data from distinct yet often complementary molecular modalities can potentially reveal novel associations that would otherwise not emerge from a piece-meal analysis of the individual datasets, and allow regulatory and disease mechanisms to be explored. For example, the functional consequences of susceptibility loci are often not well understood, but by combining SNP data with expression data, the causal relationships between the molecular and phenotypic traits of interest can begin to be explored [12]. Similarly, the role of CpG methylation in gene regulation can be better understood by including gene expression data [13]. There is great demand for statistically rigorous and computationally efficient methods to achieve such aims, and although many tools already exist, information is sparse and difficult to find, and wide-spread adoption of such methods has been limited.

5.1.3 *Classifying vertical integrative methods*

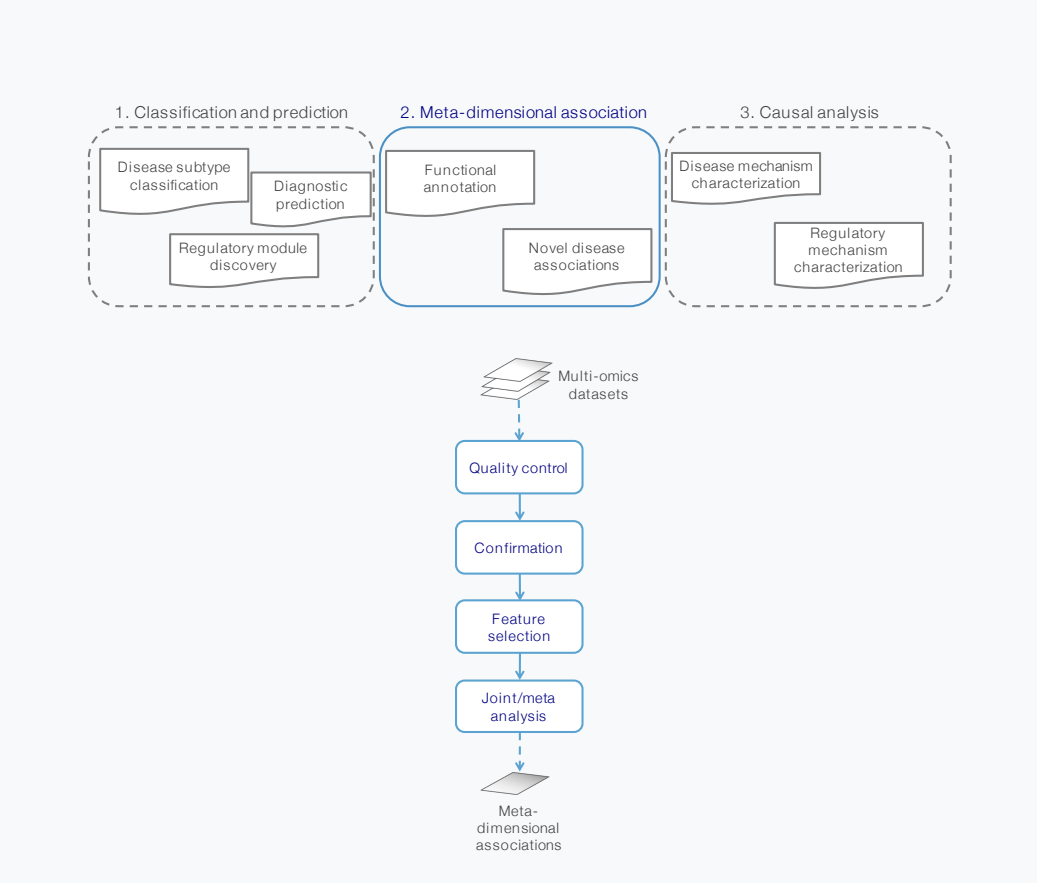

Figure 5.1. An overview of vertical integration

Here, we identify the three main classes of vertical integrative problem : 1. Classification and prediction - for disease subtype classification, regulatory module discovery, and diagnostic prediction, 2. Meta-dimensional association - for functional annotation and identification of novel disease associations, 3. Causal analysis - for the characterization of molecular mechanisms involved in disease or gene regulation (Figure 5.1). In the context of this thesis, we are specifically interested in meta-dimensional association, as a means to utilise the available gene expression and methylation datasets available for our sample of MZ ASD twins. We propose a analysis flow with the following stages: 1) quality control, 2) confirmation, 3)

feature selection, 4) joint/meta analysis. Strategies for the other two classes of integrative problem will not be discussed as firstly we do not intend to address these particular biological questions in our study, and secondly, as they already achieve a great deal of attention in the statistics and bioinformatics literature. For classification and prediction, we direct the reader to a recent review covering mathematical aspects of predominantly classification type problems in an integrative context [14], and two more reviews discussing the particular challenges involved and methods used for integrative classification in the context of cancer genomics [15] [16]. For causal analysis, we direct the reader to a paper by Schadt et al. describing a multi-step filtering approach followed by Bayesian network reconstruction and model selection to identify the most likely causal model describing the data [17], and a paper by Relton et al. describing the technique of Mendelian randomisation applied to epigenetic epidemiology, using genetic data to infer the direction of causality between epigenetic differences and the trait under consideration [18].

5.1.4 Analytical challenges and considerations

Integrating diverse, complex omics data presents a variety of analytical challenges, as discussed in [7, 13]. The issue of high data dimensionality is particularly relevant, as due to the typical sample sizes used in omics studies, there are likely to be much fewer subjects than parameters (n « p problem). High dimensionality can severely limit statistical power, making the threshold for significance of any associations difficult to overcome. In addition, having many dimensions can also be a computational burden, especially for an activity such as model selection, where for example, to calculate all possible models for three variables would require the testing of approximately 2×10^9 different models [7]. As a result, many integrative methods necessarily include dimensionality reduction as a step.

Another challenge associated with data integration is heterogeneity between the datasets. Data from different sources are measured on different scales and are likely to have different distributional properties. For example, measurements of gene expression from an RNA-seq experiment can have integer values ranging from 0 to

1000s, with approximate negative binomial distribution, whereas methylation as profiled by microarray is usually provided in the form of Beta-values, which can have a value from 0 to 1 with a bimodal distribution. There is also the issue of different omics sets taking measurements over different genomic scales or units, for example single base, exon, gene or haplotype block. A simple approach would be to aggregate measurements to bring them on to a common scale, but this is perhaps of questionable biological plausibility and also then limits the analysis to investigating *cis* effects. Another option could be to identify multiple features of interest to test across the sets, but this then increases the multiple testing burden. Finally, there are also dataset specific considerations such as the presence of confounding and batch effects, which need to be adjusted for. These various issues can complicate statistical modelling, preventing datasets from being incorporated into the same regression model for example, or making the search-space unfeasibly large for model selection.

Multicollinearity between datasets can also be an issue, with features likely to be correlated between sets. This can create difficulties further downstream, for example in regression-based approaches, where a high level of correlation might prevent matrix inversion, required for accurate coefficient estimation [19].

Finally, there is the question of how to interpret potentially complex findings. The identification of susceptibility loci, even with the added level of information that other datasets can add, is still difficult to relate to the outcome mechanistically. This is perhaps where a systems-level view placing the identified perturbations within the context of interaction networks, might prove more fruitful [12].

5.1.5 Integrating multi-dimensional omics datasets

This chapter is divided into two sections. In the first section, for each of the stages of the proposed meta-dimensional association analysis flow, some theoretical background is covered and existing bioinformatics tools highlighted. In the second section, a number of selected approaches are applied to biological basis of autism spectrum disorder (ASD). In this case study we attempt to integrate expression and

methylation datasets from the same sample of MZ twins concordant and discordant for ASD, in order to maximise power to identify functionally relevant signals across the datasets and help form a more coherent picture of the genes and mechanisms likely to be involved in the aetiology of the disorder.

5.2 Meta-dimensional analysis

5.2.1 Quality control

Prior to combining evidence, assurances will often be sought that the individual datasets are of a high quality. Each profiling technology generates data with its own characteristics and has its own recommendations with regards to quality assurance and quality control, and adhering to these standards should help to reduce the risk of encountering issues further downstream [7].

5.2.2 Confirmation

The aim of confirmation is to establish the extent to which the measurements from the diverse modalities are related. If the study aims to explore global molecular differences between subjects, comparing correlation across all features measured in common across the datasets can be useful for gaining an impression of the overall agreement between the datasets. Alternatively, individual gene-wise correlations are informative about functional relationships between different molecular measurements at individual loci. For example, by testing the correlation between the abundance of a gene and the methylation status of a *cis* CpG site allows for the identification of methylation-expression quantitative trait loci (methyl-QTL) - genes whose expression is under the control of a *cis* CpG.

In terms of the statistical methods, classic approaches such as Pearson's and Spearman's correlation are often appropriate. A number of studies have integrated methylation and gene expression data using such approaches. For example Rhee et al.

[20] used correlation as part of a multi-staged integrative analysis in breast cancer, first identifying differentially expressed genes for cancer subtype, and then examining the correlation between methylation and expression at each locus, verifying the functional impact of DNA methylation in gene expression in relevant genes. In another study, Zhao et al. integrated mRNA expression with both methylation and miRNA to identify differentially regulated genes in CD4+T cels from systemic lupus erythematosus patients [21].

Only limited inferences can be made using such methods, firstly due to the fact that only two datasets can be correlated at a time, and secondly, because it is not typically possible to determine the direction of the effect. As a result, this type of analysis is often performed as a purely exploratory step, or to filter data before performing another type of integration. For example, correlating SNP and expression data to identify likely eQTLs could be used to filter features for subsequent modelling along with the trait.

5.2.3 Feature selection

One of the challenges of data integration is high dimensionality, as discussed previously. Reducing the number of variables being tested for association is an important part of an integrative analysis - in fact some of the individual methods for joint analysis include an inherent feature selection or dimension reduction step (e.g. penalized regression - see next section). Feature selection or dimension reduction is important in the context of classification and prediction, where it can be used as an inferential method in its own right, but for meta-dimensional association, it could perhaps be considered more of an exploratory step used to filter the data before performing inference - especially when the inferential method does not include an inherent dimension reduction step (e.g. standard regression).

As for existing methods, machine learning methods such as principal components analysis (PCA) or canonical correlation analysis (CCA) are often used. CCA works similarly to PCA describing associations between datasets by finding linear combinations that maximise the correlations between the sets, but whereas PCA is pre-

dominantly used to describe variance in a single set, CCA can be used for two datasets [22]. The PMA package in R (https://cran.r-project.org/web/packages/PMA/index.html) can be used to perform both sparse CCA and PCA [23]. Alternatively, the mixOmics R package (https://cran.r-project.org/web/packages/mixOmics/index.html) provides regularised CCA designed to deal with a large number of variables as well as a sparse partial least squares method (PLS) [24] .

5.2.4 Joint analysis

5.2.4.1 Regression

Linear regression and similar methods are popular for performing statistical tests with single omics data. Intuitively, such methods could also be used in an integrative setting where the goal is to identify meta-dimensional associations. These potentially offer several advantages over alternative meta-analytic approaches (discussed in the next section). When the genomic datasets are particularly rich in information for example, the effect of extraneous or confounding variables can perhaps be more accurately estimated by combining the datasets. In cases too where a causal hypothesis is to be tested, regression models are more amenable to testing for directionality of an effect - this however would typically require the application of another method such as model selection and/or network reconstruction.

Once data has been appropriately transformed, scaled to be approximately normally distributed, and genetic features brought on to the same scale, in the simplest case where two datasets are to be combined, a multivariate or multivariable model can be fit. For example, using SNP and methylation data, methylation is treated as an intermediate trait and regressed on genotype groups. Bell et al. used such an approach, summarising methylation by haplotype block and using sliding windows to investigate the relationship between genotype and methylation for the FTO diabetes susceptibility locus [25].

There are a number of R and bioconductor packages which use regression as the

core method for integrating two datasets. MethylMix (https://www.bioconductor.org/packages/ release/bioc/html/MethylMix.html) is intended for integrating methylation and expression data to determine significantly hypo and hypermethylated transcriptionally predictive genes in cancer [26]. The MEAL package (http:// bioconductor.org/ packages/MEAL/) can also integrate methylation and expression data using a linear regression-based approach, as well as containing several methods for differential methylation analysis and probe/region summarization.

Regression with dimension reduction: When the number of molecular datasets to be integrated is > 2, and the number of variables is much greater than observations, traditional multivariate approaches (such as those described above) are no longer appropriate [27]. In such instances, feature selection or dimension reduction methods can instead be used. Methods such as partial least squares (PLS), Least Absolute Shrinkage and Selection Operator (LASSO) [28], and Elastic Net (ENET) [29], combine dimension reduction with regression, removing uninformative parameters and hence mitigating issues with dimensionality. Additionaly, these methods can deal with multicollinearity [30], which standard regression methods are less able to deal with.

The plsgenomics R package (https://cran.r-project.org/web/ packages/plsgenomics/index.html) implements PLS spefically for integrative omics applications, its utility demonstrated for combining expression and ChiP data to predict transcription factor activities in the accompanying publication [31].

5.2.5 Meta-analysis

5.2.5.1 Combining evidence

As an alternative to regression based approaches, data can instead be combined at the level of summary statistics such as odds ratios, test statistics or p-value. This approach has a long history in meta-analysis [32], and more recently p-value combi-

nation methods have found application in GWAS [33], multiple testing correction [34], and geneset testing methods [35]. Such methods could also find application in vertical data integration. The main advantages are that the raw experimental data is not always required, making it suitable for including extraneous data sources (e.g. annotation, or publicly available GWAS summary data), and also circumventing potential issues with data heterogeneity. The real value in such an approach is that it is possible to gain statistical power, even the case where evidence from the individual studies is weak [36]. Yet, despite the potential advantages, the idea of using horizontal integration techniques for vertical integration has not been explored much in the literature.

In choosing an evidence combining approach, several important factors relating to the characteristics of the datasets and also the type of combined hypothesis being tested should be carefully considered. Related to this, whether the measurements are derived from the same individuals or unrelated individuals, and if the datasets are fully independent or if, on the other hand, the measurements across the datasets are correlated. As for the nature of combined test, the majority of methods test the global null, sometimes known as the conjunction test, which tests for at least one of null hypotheses being false, but different combined hypotheses can also be tested (see below). Lastly, it is important to consider whether the strength of evidence from each study should be weighted equally - study weights are more easily incorporated into some approaches than others.

5.2.5.2 Methods for combining independent p-values

Quantile combination - Fisher's and Stouffer-Liptak: Methods for combining p-values can broadly be categorised into those based on either quantile combination or order statistics. Quantile combination methods first select a parametric cumulative distribution and then transform p-values into distributional quantiles through a combining function $\Psi = \sum_{i=1}^{k} f(p_i)$, where f is the transforming function and p_i the i^{th} p-value from k tests, and use this to calculate a combined test p-value from the sampling distribution of Ψ [37]. Probably the best known quantile-based approach is Fisher's combined p-value, which models combined p-values as a χ^2 dis-

tribution using the following transformation: $\Psi_F = -2 \sum logp_i$. One limitation of Fisher's method is that it cannot incorporate weights, an issue particularly in meta-analysis where it may be desirable to weigh statistics according to the reliability of the study, or perhaps by how informative it is judged to be [38]. The Stouffer-Liptak method (equivalent methods developed independently) is an extension to Fisher's which allows for weights to be incorporated by instead combining Z scores with: $\Psi_{SL} \sim \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$. Or in terms of p: $\Psi_{SL} = \sum \Phi^{-}1(1 - p_i)$ where Φ^{-1} is the inverse cumulative distribution function of the normal distribution.

Order statistics - Tippet's and rank truncated product: Order statistic methods use a combining function: $\Psi = p_r$ for a preselected $r, 1 <= r <= k$, with the sampling distribution a member of the beta distribution family. Tippet's method simply takes the minimum p-value and uses this as the combined value: $\Psi_T = min(p_i)$. There are also more recently developed methods such as rank truncated product and adaptive rank truncated product which could perhaps be considered hybrid approaches, making use of both order statistics and non-parametric cumulative distribution modelling. In the case of adaptive rank truncated product minimum p-values are found to be accurately modelled by beta distribution, which can be used to assess significance (see Chapter 4).

The distinctions between the different combination approaches will not be discussed any further here (see [37] for a review). Much more relevant are their similarities in terms of the underlying assumptions made. Firstly, these combination methods rely on the fact that for test statistics having continuous null distributions, the p-values will be uniformly distributed across the interval 0 to 1 [37]. Secondly, the conjunction, global, or combined null hypothesis is tested, where the alternative is that at least one of the null hypotheses is false. Finally, individual p-values are assumed to be fully independent, that is to say, the measurements are uncorrelated within and between assays.

P-value combination in multi-omics analysis: One of the difficulties with applying the above classic combination strategies to multi-omics data, is that some of

the assumptions of the methods are violated. The issue of dependence in particular is problematic, given for example, well described phenomena in genetics such as linkage disequilibrium (LD), whereby markers show correlation as a result of being inherited non-randomly in haplotype blocks. As well as individual datasets possessing their own complex correlation structures, when using different molecular datasets derived from the same sample, there can also be correlation between datasets. Applying Fisher's method with correlated data has been shown in simulations to produce unexpectedly small p-values, thus over estimating the significance of the combined hypothesis and increasing the Type I error rate of the test [39].

5.2.5.3 Combining dependent p-values

Brown's, empirical Brown's and Kost's: Brown was one of the first to suggest a method for combining dependent p-values. Brown's method is an extension of Fisher's, in which p-values are assumed to come from a multivariate normal distribution with specified covariance matrix [39]. A scaled χ^2 distribution is used: $\Psi_B \sim c\chi^2_{2f}$, where the constants f and c represent the re-scaled degrees of freedom and a scale factor which is the ratio between the degrees of freedom in Fisher's and Brown's [40]. These are calculated as follows:

$f = \frac{E[\Psi]^2}{var[\Psi]}$

$c = \frac{var[\Psi]}{2E[\Psi]} = \frac{k}{f}$

$E[\Psi] = 2k$

$var[\Psi] = 4k + 2\sum_{i<j} cov(-2logP_i - 2logP_j)$

$\Psi_B = 1 - \phi_{2f}(\Psi_F/c)$

where $\phi 2f$ is the CDF of χ^2_{2f} and Ψ_F is the distribution for Fisher's method (as previously defined).

Because the expectation and variance of Ψ_B are calculated directly via multidimensional numeric integration, the method is impractical for large datasets [40].

This is a clearly an important issue in genomics, and has consequently motivated the development of methods based on Brown's, which use approximations of the covariance term to reduce the computational burden. The first of these is a method described by Kost et al. [39], in which exact co-variances for Fisher's distribution were computed across a grid of different correlation (ρ) values and differing degrees of freedom (v), and different polynomial regression models fitted. Approximations for covariance as a function of ρ and $1/v$ were then obtained, in both cases where variance is known and unknown, these can then used by substituting into Brown's (formulae not given here for the sake of brevity). The second method is the empirical Brown's approach of Poole et al. [40], which similarly uses approximation, but achieves this through alternative means by deriving an empirical calculation for the covariance term in Brown's from the experimental data itself:

$\vec{w}_i = -2logF(\vec{x}_i)$

$var[\Psi] = 4k + 2\sum_{i<j} cov(\vec{w}_i, \vec{w}_j)$

where $\vec{w}_i$ is a transformed vector based on $\vec{x}_i$, samples from the raw data, and the empirical CDF calculated from the sample $\vec{x}_i$. The covariance is then calculated using $\vec{w}_i$. Both Kost's and empirical Brown's methods are implemented in the EBM bioconductor package
https://www.bioconductor.org/packages/ devel/bioc/html/EmpiricalBrownsMethod.html

Stouffer-Liptak-Kechris: As mentioned above, another issue with particular relevance to omics data is that of correlation. This has implications for combining not only genetic data, but also methylomic data too, with the methylation status of CpGs showing patterns of correlation for distances up to several thousand base pairs (see Chapter 4). The Stouffer-Liptak-Kechris (SLK) method, uses sliding window correction where neighbouring p-values are weighted according to the observed auto-correlation, before undergoing combination by Stouffer-Liptak [41]. The SLK method has been implemented in the Comb-p tool in python (https://github.com/brentp/combined-pvalues), which uses p-combination to identify "peaks", or genomic regions of statistical significance within an individual dataset, which can be applied to methylomic, or in theory any molecular dataset with inherent

spatial correlation [42] . Although it is not explicitly mentioned in the paper, SLK could also potentially be used in multi-omics integration, as a method to first account for correlation within-dataset, as a precursor to then integrating across datasets. This proposed framework for meta-dimensional association involving multiple steps of p-value combination is explored further in the second section of this paper.

5.2.5.4 Testing the partial conjunction

Touching briefly on one last aspect of p-combination which warrants consideration, is the question of which hypothesis to test. All of the methods mentioned previously test the conjunction hypothesis. With multiple omics datasets, conceivably, testing that just one of the alternative hypotheses is true might not be biologically meaningful, or adequately reflect the underlying molecular mechanisms. Using an example to illustrate this point, in an investigation including SNP, methylation, and CHIP-seq datasets, for any individual susceptibility locus, ideally we would like evidence for association across all the datasets, not just one. The disjunction hypothesis, which tests that all of the alternative hypotheses are significant, would be overly restrictive, particularly since the relationship between methylation status of individual CpGs or regions and chromatin status is not well understood - potentially operating in either direction, distally, or indeed having no effect at all.

One solution to this scenario would then be to use the partial-conjunction, which tests that a pre-specified n number of alternative hypotheses at a location are true [36]. To use another example, when combining SNP, DNA methylation and gene expression data, we might specify $n \geq 2$, and perhaps further require that one of the significant p-values is a genetic association, to increase the likelihood that any identified signal will be a functionally relevant eQTL or mQTL. By relaxing the constraint for the other two functional datasets by only requiring one show an association, we are then able to screen for *cis* mQTLs, eQTLs, methQTLs, and not just methQTLs (as we would be doing with the disjunction). Benjamini and Heller [36] describe appropriate test statistics for a partial conjunction, the Matlab code is available from the author's website (http://www.math.tau.ac.il/ ruheller/Software.html).

5.2.5.5 *Utility of p-value combination in vertical integration*

Within an integrative analysis framework p-value combination methods can be used to achieve a couple of different analytical aims. Firstly, p-value combination can be used for data filtering prior to performing a modelling-based integrative approach. The results from p-value combination can also be used stand-alone, where significant meta-dimensional association indicate a dependence relationship between the molecular traits and phenotypic trait of interest, although determining the direction of the relationship may require the application of additional methods.

5.3 Application of selected integrative methods to ASD

This section describes the application of some of the above methods to integrate 27K microarray methylation data and RNA-seq gene expression for our ASD MZ twin cohort, in order to identify meta-dimensional associations and highlight functionally relevant signals. Further information on the sample is given in Chapters 2 and 3. The 27K methylation data was generated as part of a previous study [43], and the data provided for this in the form of pre-processed and normalized Beta values, see the original paper for details on the methods used and data processing steps. The RNA-seq data was generated as previously described in Chapter 3, and the final processed data in the form of normalized gene counts per million (NCPM) used here along with the differential expression results.

5.3.1 *Methods*

5.3.1.1 *Quality control*

To begin with, the expression and methylation datasets were compared, and any individual not profiled on both assays removed. Following this step, the methylation

Beta values were converted to M values (see Chapter 4) and differential methylation analysis performed. The ComBat function from the SVA package [44] was used to identify potential confounders and other unmeasured sources of variation, which in this case failed to identify any significant surrogate variables. PCA plots were also generated (see Appendix D Figure D1) and inspected, with no evidence of batch effects or confounding observed. Next, edgeR was used to fit a linear model for a case-control comparison and a within group 2 comparison, using the same model that was used for the expression data (see Chapters 3 and 4):

Case-control:

$$y \sim pair + case \tag{5.1}$$

Group 2 pairs:

$$y \sim pair \tag{5.2}$$

The results from differential methylation analysis were compared with those published in the original research, the results compared, and top results showing good overlap.

5.3.1.2 Confirmation

For confirmatory integration, firstly, Pearson's correlation between samples across the assays was assessed using Beta values for the methylation data and NCPM for the expression data. Probes were binned according to genomic location - all, CpG island, non-CpG island. Mean values for all the samples and CpG types were calculated, and 95% confidence intervals calculated using Fisher's z transformed values. Following this, the correlation and intersection between the differential methylation and differential expression results were measured using the reported log fold change values for all genes and the significant genes at $p < 0.05$. Finally, Pearson's correlation between genes was assessed, by aggregating CpG probes and taking the median Beta value for the gene.

5.3.1.3 *Joint analysis - p-value combination*

To combine evidence across the sets, two rounds of p-value combination were used. In the first, within-set combination step, in order to account for correlation within the 27K dataset, the SLK method from comb-p was used setting the distance parameter to 2000 and region filter to 0.05. This produced SLK adjusted p-values for each CpG site. Following this, sites mapping to the same gene in the 27K data were aggregated by taking the minimum p-value as the gene-wise significance estimate. The gene expression data was assumed not to have any spatial correlation, and so was not subject to any initial round of combination. In the second step, to combine the SLK-adjusted methylation p-values and gene expression p-values, empirical Brown's was used to obtain the final combined p-value for genes measured in both assays, using the original raw data in the form of NCPM values for the expression and M values for the methylation. In cases where a gene was unmeasured in either assay a p-value of 1 was assigned to that missing gene, and since the covariance calculation would in that case not be possible, Fisher's method was instead used - thus ensuring that a significant result from a single dataset that unmeasured in the other assay would still remain significant, in order to produce a complete combined set of results.

5.3.1.4 *Joint analysis - regression modelling*

Next, a regression-based model selection approach was used to combine the data within the same model. The data was filtered to retain only loci which were measured in both expression and methylation assays that achieved nominal significance in the previous p-value combination approach. Next, the expression NCPM were transformed to Z scores in order to bring them onto a smaller scale closer to that of the methylation M values, and the following three mixed effects models fitted:

Model 1:

$$logit(case) \sim 1|pair + methylation \tag{5.3}$$

Model 2:

$$logit(case) \sim 1|pair + expression \tag{5.4}$$

Model 3:

$$logit(case) \sim 1|pair + methylation + expression + methylation : expression \tag{5.5}$$

With 1 | pair the random effect of pair, and alternatively fixed effects for methylation, expression, and the interaction between methylation and expression in models 1,2, and 3 respectively. For each gene tested, the Akaike information criterion (AIC) was calculated for each of the models using the *AIC* function in *R*. The AIC is calculated as : $AIC = -2\times(L)+2k$, where the L is the maximum log-likelihood for the model and k is the number of parameters. This favours models with a good fit but penalizes those with more parameters (to prevent overfitting). In cases where model 3 provided the lowest AIC score, this indicated that the inclusion of methylation and expression and their interaction improved prediction of case status even at the expense of increasing model complexity. This was used to identify functionally relevant signals, defined as those genes that in addition to showing significant association with ASD in either the expression or methylation datasets, also showed evidence of association between methylation status and expression level.

5.3.2 Results

5.3.2.1 Confirmation

The per-sample interset correlation was assessed by comparing per-gene Beta values to gene expression NCPM. The results are given in Table 5.1. Samples showed a modest level of correlation across the datasets, with a mean value of -0.15, 95% CI[-0.49,0.18]. This was slightly lower when only taking into account CpGs located in islands -0.13, 95% CI[-0.46,0.21], and higher for non-island CpGs -0.21, 95% CI[-0.55,0.12]. These results were taken as indicative of a relationship between the datasets, even though the correlations would perhaps be expected to be slightly

higher.

The agreement of the results from differential methylation and differential expression analyses was assessed by correlation and intersection between the mean log fold change (logFC) values for all genes, and nominally significant genes. The results are shown in Table 5.2. For the case-control comparison, the correlation between the logFC values for all genes was -0.05, and -0.20 for significant genes. For the group 2 comparison, the correlation between the logFC values for all genes was -0.05, and 0.02 for significant genes. Based on these values, it was decided that further integrative analysis should utilise only the case-control results, as these would be more likely to reveal novel meta-dimensional associations - given the higher correlation between significant genes across the platforms.

Overall correlation between gene measurements across the different assays was also examined. Table 5.3 gives the top 50 genes in terms of absolute correlation of per-gene mean CpG methylation Beta values and expression NCPM.

Comparison	**Sample correlation (p)**	**95% CI lower**	**95% CI upper**
Betas (All CpGs) with CPM	-0.15	-0.49	0.18
Betas (Island CpGs) with CPM	-0.13	-0.46	0.21
Betas (Non-island CpGs) with CPM	-0.21	-0.55	0.12

Table 5.1. The per-sample interset correlation between mean CpG methylation (Beta) and transcript abundance (normalised counts per million) across all genes, with CpGs classified by location (all, island, non-island)

	logFC all genes		**logFC DE genes (P < 0.05)**	
Comparison	*Correlation (r)*	*Intersection*	*Correlation*	*Intersection*
case control	-0.05	0.55	-0.20	0.06
group 2	-0.05	0.55	0.02	0.04
gender	0.00	0.55	-0.13	0.11

Table 5.2. Agreement of the results from differential methylation and differential expression analyses, as assessed by correlation and intersection between the mean log fold change (logFC) values for all genes, and nominally significant genes.

Gene symbol	Correlation (r)
LDHC	-0.851865076
IL1R2	-0.806544669
PNMA3	0.797076607
IL5RA	-0.758055223
BTN3A2	-0.735685641
FADS2	-0.731084493
IRF6	-0.720178734
TREM1	-0.706696357
UBE4A	0.682398902
LRG1	-0.678783238
C8orf31	-0.668632019
FUT7	-0.666878672
HLA-C	-0.662617802
RAB11FIP2	0.655762494
CSF3R	-0.65400064
USP10	-0.652070938
ZNF691	0.645900146
GNLY	-0.644868421
TRPM6	-0.643086325
CD3G	-0.63391781
ATG10	0.629443839
GZMM	-0.609802
NFE2	-0.607621139
HIPK3	-0.602083315
ENTPD1	-0.600549279
AQP9	-0.600356036
HAL	-0.59864147
CXADR	-0.598117028
SCNN1D	0.591966877
DDX43	-0.590682908
ATP13A4	0.588722496
CD19	-0.588227811
TREML2	-0.583088408
FBXL13	-0.581662955
EBPL	-0.57808595
IL1B	-0.577854054
KRT1	-0.569225662
CEACAM3	-0.564862165
MKRN3	-0.564686395
ZNF205	-0.564236741
CHI3L1	-0.563905026
PRSS21	-0.562284555
WDR45	0.557622035
SIGLEC11	-0.556642295
ARMC1	-0.554141787
ZFX	-0.553541715
ITGAX	-0.55309672
RNF17	-0.552319755
LY9	-0.550353001
RYBP	-0.542635604

Table 5.3. The top 50 genes in terms of absolute correlation of per-gene mean CpG methylation (Beta) and transcript abundance (normalised counts per million)

5.3.2.2 Joint/meta analysis - p-value combination

Turning now to the p-value combination method, Table 5.4 gives the top 50 results of p-combination for the case-control comparison. Here it is apparent that a number of the genes previously identified as being differentially expressed in the RNA-seq experiment continue to be highly significant after combination, making up the majority of the top 10. One example is *DEPDC1B*, which is not significantly differentially methylated, but has enough evidence for differential expression so that it remains significant. Table 5.5 gives the top 50 results of p-combination for the group 2 comparison. Once again, the signal from the differentially expressed genes seems to overwhelm that of the differentially methylated genes. Potential reasons for this are numerous and could arise from inherent differences in the molecular measurements, for example the different scales of the units being measured and the need to combine CpG measurements in order to obtain gene-level estimates (which may not reflect the underlying regulatory mechanisms), or perhaps in the expected effect sizes and power to detect these.

Case-control

Gene symbol	27K DM (SLK adjusted p)	RNA-seq DE (p)	Fisher's/E-brown's combined (p)
ZNF501	5.56E-01	2.96E-05	2.15E-04
IGHG4	1.00E+00	2.86E-05	3.28E-04
DEPDC1B	9.67E-01	2.05E-05	3.37E-04
ALKBH6	4.88E-02	1.14E-03	6.01E-04
PHKA2	2.16E-03	2.58E-02	6.36E-04
PRR13P5	1.00E+00	6.00E-05	6.43E-04
IDUA	5.67E-01	1.36E-04	8.09E-04
KBTBD8	1.59E-02	2.10E-03	9.81E-04
PARP6	6.31E-02	1.60E-03	1.05E-03
ARHGEF1	3.43E-03	4.40E-02	1.48E-03
MOCS3	2.96E-02	6.16E-03	1.75E-03
HOXA9	1.88E-04	1.00E+00	1.80E-03
C11orf49	2.13E-02	8.89E-03	1.81E-03
TINF2	8.77E-04	2.16E-01	1.82E-03
RNF43	1.04E-02	1.79E-02	1.82E-03
TAS2R60	4.24E-03	2.67E-02	1.83E-03
IMPA2	4.95E-04	3.90E-01	1.84E-03
ZNF499	1.96E-04	1.00E+00	1.87E-03
HCFC1R1	1.73E-02	7.49E-03	1.87E-03
AMPD3	8.29E-03	1.65E-02	1.87E-03
ASPM	1.18E-02	1.35E-02	2.05E-03
DNAJA1	8.43E-03	2.60E-02	2.07E-03
SNRPE	7.78E-02	3.06E-03	2.22E-03
ANXA1	8.44E-04	1.54E-01	2.24E-03
ANP32E	3.00E-01	3.68E-04	2.26E-03
PFKP	2.54E-04	9.80E-01	2.32E-03
MCM4	3.95E-03	4.51E-02	2.32E-03
RECQL5	1.76E-03	1.46E-01	2.37E-03
GADD45GIP1	3.01E-01	8.94E-04	2.48E-03
CKB	2.04E-02	9.23E-03	2.60E-03
BRIP1	1.85E-02	1.56E-02	2.64E-03
CDK15	1.00E+00	2.89E-04	2.64E-03
PPP1CC	9.11E-03	2.82E-02	2.65E-03
TTC7B	1.00E+00	3.10E-04	2.81E-03
MAP4	7.40E-03	4.29E-02	2.88E-03
RHOBTB1	6.45E-02	1.80E-03	2.89E-03
SSNA1	4.91E-02	6.88E-03	3.04E-03
GABBR1	3.18E-02	8.68E-03	3.21E-03
SPATA6	1.59E-01	2.30E-03	3.25E-03
PLAGL1	4.99E-02	5.95E-03	3.33E-03
MAPK8IP3	2.59E-02	9.04E-03	3.38E-03
PABPC1L	1.00E+00	3.86E-04	3.42E-03
CDC40	1.33E-02	3.04E-02	3.57E-03
STK11IP	1.00E+00	4.12E-04	3.62E-03
APOA1BP	1.93E-03	2.20E-01	3.73E-03
HSPA8P14	1.00E+00	4.33E-04	3.79E-03
C21orf62-AS1	1.00E+00	4.38E-04	3.82E-03
ASF1A	2.72E-02	1.62E-02	3.84E-03
SERBP1	3.42E-03	1.30E-01	3.87E-03
MPL	7.33E-02	4.10E-03	3.94E-03

Table 5.4. Top 50 genes for the case-control comparison following p-value combination using SLK to summarise the methylation data, and then either Fisher's or empirical Brown's method to combine with expression data, depending on whether the gene was measured in both assays.

Group 2

Gene symbol	27K DM (SLK adjusted p)	RNA-seq DE (p)	Fisher's/E-brown's combined (p)
IGHG4	1.00E+00	1.50E-08	2.86E-07
SNORD15B	1.00E+00	6.13E-06	7.97E-05
EVI2A	4.28E-01	5.25E-06	9.59E-05
ASPM	6.35E-03	2.19E-03	2.54E-04
DNAJA1	8.61E-03	2.52E-03	2.55E-04
CEP55	1.13E-01	1.93E-04	2.69E-04
RGS18	1.00E+00	3.39E-05	3.83E-04
RPS3A	2.56E-01	1.60E-04	4.55E-04
LPAR6	1.00E+00	5.09E-05	5.54E-04
RPL9	5.83E-01	6.16E-05	6.05E-04
HOXA9	7.73E-05	1.00E+00	8.09E-04
AMPD3	2.12E-03	3.04E-02	9.91E-04
PRR13P5	1.00E+00	1.00E-04	1.02E-03
RNF43	4.07E-03	2.64E-02	1.12E-03
RPL21	2.21E-01	5.25E-04	1.17E-03
PPBP	1.00E+00	1.37E-04	1.35E-03
ANXA1	3.66E-04	1.91E-01	1.36E-03
S100A12	2.42E-01	5.71E-04	1.37E-03
RECQL5	6.11E-04	2.38E-01	1.43E-03
DAPP1	1.69E-03	9.05E-02	1.50E-03
TAS2R60	3.95E-03	2.27E-02	1.51E-03
KBTBD8	2.38E-02	2.51E-03	1.53E-03
TPT1	1.71E-02	6.57E-03	1.56E-03
B2M	6.51E-01	2.58E-04	1.63E-03
IDUA	6.04E-01	2.90E-04	1.69E-03
DEPDC1B	9.65E-01	1.37E-04	1.71E-03
TINF2	5.50E-04	3.31E-01	1.75E-03
EEF1A1	2.58E-01	5.93E-04	1.89E-03
RPL41	1.00E+00	2.04E-04	1.94E-03
HSPA8P14	1.00E+00	2.20E-04	2.07E-03
HCFC1R1	2.92E-02	5.42E-03	2.20E-03
CDC40	4.92E-03	4.87E-02	2.24E-03
ALKBH6	4.25E-02	5.85E-03	2.31E-03
MKRN3	6.67E-04	3.76E-01	2.33E-03
RPL30	1.00E+00	2.53E-04	2.35E-03
MMP8	2.65E-02	9.79E-03	2.40E-03
TTC7B	1.00E+00	2.70E-04	2.49E-03
COPS4	3.38E-02	8.10E-03	2.52E-03
SRGN	1.00E+00	2.86E-04	2.62E-03
HIST1H2AG	1.47E-01	2.00E-03	2.68E-03
ZNF501	5.68E-01	4.93E-04	2.71E-03
MOCS3	1.68E-02	1.79E-02	2.73E-03
C11orf49	1.08E-02	2.88E-02	2.82E-03
RPL39	2.45E-01	1.33E-03	2.94E-03
HIST1H3J	5.89E-01	3.04E-04	2.95E-03
PHKA2	6.46E-03	4.90E-02	2.98E-03
PPP1CC	4.11E-03	7.66E-02	3.16E-03
RPL23	2.35E-02	1.01E-02	3.22E-03
GADD45GIP1	2.12E-01	1.74E-03	3.28E-03
RPS15A	3.34E-01	1.11E-03	3.30E-03

Table 5.5. Top 50 genes for the within group 2 comparison following p-value combination using SLK to summarise the methylation data, and then either Fisher's or empirical Brown's method to combine with expression data, depending on whether the gene was measured in both assays.

5.3.2.3 Joint/meta analysis - regression

Finally, the significant results from p-value combination were used to select genes for model fitting. Those genes where the interaction model provided the best fit to the data were retained and then annotated with correlation values and combined p values from the previous steps. A final list of functionally relevant genes was generated, defined as those with a significant combined p-value ($p < 0.05$) and significant correlation ($-0.2 < r > 0.2$). This produced a list of 45 potentially functionally relevant genes. The results are shown in Table 5.6 and discussed below.

gene	corr	pval	slk.P.adj	expr.P	fishers.browns.P
TINF2	-0.25	2.00E-16	8.77E-04	2.16E-01	1.82E-03
SNRPE	-0.26	2.00E-16	7.78E-02	3.06E-03	2.22E-03
RHOBTB1	0.28	3.98E-01	6.45E-02	1.80E-03	2.89E-03
ASF1A	-0.45	2.00E-16	2.72E-02	1.62E-02	3.84E-03
SLC7A6	-0.50	8.93E-01	3.77E-02	1.66E-02	5.23E-03
SF3B2	0.33	4.10E-12	3.33E-02	1.22E-02	5.77E-03
MAGED1	-0.21	1.30E-01	7.61E-03	1.03E-01	6.38E-03
TTF2	0.30	1.77E-02	2.65E-03	1.46E-01	6.90E-03
RAP2A	-0.28	1.91E-27	1.55E-01	5.97E-03	7.38E-03
PDK2	-0.27	2.00E-16	9.15E-01	1.41E-03	9.88E-03
PRKCA	0.20	7.33E-01	5.19E-02	1.98E-02	1.01E-02
RHOF	-0.21	2.78E-01	2.02E-02	6.62E-02	1.02E-02
KCNE1	-0.35	5.40E-01	2.92E-01	4.77E-03	1.05E-02
HMBOX1	-0.26	2.00E-16	9.10E-02	1.58E-02	1.08E-02
CAMK2G	-0.23	1.01E-01	6.95E-03	2.15E-01	1.12E-02
MC1R	0.24	5.48E-01	1.08E-02	1.32E-01	1.13E-02
PCSK7	0.31	3.14E-02	7.66E-03	2.04E-01	1.17E-02
BRD3	0.48	2.00E-16	4.24E-02	2.65E-02	1.41E-02
ZDHHC24	-0.26	2.07E-01	6.89E-03	3.00E-01	1.49E-02
FER	-0.24	8.33E-01	7.49E-03	3.07E-01	1.63E-02
TSPAN15	-0.22	1.77E-12	4.01E-02	5.93E-02	1.67E-02
B4GALNT3	-0.21	1.73E-01	3.92E-01	5.51E-03	1.68E-02
CRIM1	-0.28	2.47E-01	4.11E-01	6.11E-03	1.76E-02
S100A9	-0.35	6.33E-02	6.48E-02	4.00E-02	1.80E-02
STIM1	0.23	2.00E-16	4.09E-02	4.28E-02	2.08E-02
XYLT1	-0.21	2.00E-16	9.26E-01	3.85E-03	2.36E-02
SEMA3B	0.23	9.29E-01	9.23E-03	3.36E-01	2.41E-02
PLSCR4	-0.21	1.63E-04	3.55E-01	1.16E-02	2.68E-02
SLC35B3	0.32	4.12E-04	1.62E-01	1.32E-02	2.87E-02
ZNF300	-0.36	6.10E-02	8.21E-01	5.61E-03	2.94E-02
MTHFD2	-0.21	2.00E-16	1.46E-01	3.38E-02	3.11E-02
UBE3A	-0.26	3.74E-01	4.79E-02	1.10E-01	3.28E-02
DYNLT3	-0.26	3.97E-01	4.36E-01	1.25E-02	3.39E-02
RPS15A	-0.26	2.98E-01	3.07E-01	1.83E-02	3.47E-02
FGFR2	-0.29	2.70E-01	5.95E-02	1.03E-01	3.73E-02
MBTPS2	0.28	2.90E-01	3.11E-02	1.35E-01	3.74E-02
TREM1	-0.71	7.22E-01	3.13E-02	2.03E-01	3.84E-02
GRIP1	0.39	1.57E-01	5.61E-01	6.04E-03	3.94E-02
TNFRSF13(	-0.21	3.29E-02	4.46E-02	1.53E-01	4.09E-02
SUMO2	0.26	1.55E-01	9.44E-02	5.73E-02	4.32E-02
TCP11L1	-0.39	7.93E-04	3.55E-01	2.13E-02	4.45E-02
TAF6L	-0.36	2.02E-01	2.66E-01	2.86E-02	4.48E-02
PFKL	0.23	9.08E-02	8.19E-03	9.54E-01	4.57E-02
NLGN2	0.27	2.20E-01	3.78E-01	1.65E-02	4.86E-02
OXTR	0.35	2.00E-16	1.19E-01	5.02E-02	4.91E-02

Table 5.6. Final list of 43 functionally relevant genes defined as those where the methylation and expression signals were significantly correlated (-0.2 < r > 0.2) and the interaction model best explained the relationship between ASD and methylation and gene expression. The results show the regression p-value, the SLK adjusted p-value for differential methylation, the p-value for differential expression, and finally the combined p-value after applying empirical Brown's method.

5.4 Discussion

In the first section of this chapter, we attempted to bring together the collective knowledge dispersed throughout the integrative genomics literature in order to understand the challenges involved and methods available for heterogenous data integration in the context of complex disease epidemiology. Three main classes of integrative problem were identified : classification and prediction, meta-dimensional association, and causal analysis. For meta-dimensional association, a suggested analysis flow was proposed featuring quality control, feature selection, confirmation, and joint analysis/meta-analysis. For each of these, some of the theoretical background was covered, and available bioinformatics tools highlighted. Particular attention was paid to meta-analysis, which has so far not received much attention in this particular setting. Such an approach might find more widespread application use as it possesses a number of advantages over joint analytic techniques. Firstly, evidence combination is firmly rooted in tried and tested statistical methods and often making use of parametric distributions, which should give some measure of confidence in the significance estimates produced. Secondly, significant interaction effects between multiple datasets can potentially be revealed, which integrative approaches which use multiple regression steps or SNP filtering could miss, due to being limited to examining two datasets at a time. Thirdly, these methods could help reduce the impact of data heterogeneity, since first combining results within datasets can be used to bring different assay measurements onto a common genomic scale - here we used the SLK combination method to derive per-gene differential methylation estimates to then combine with gene expression measurements. Also related to data heterogeneity, when performing interset combination there is no need to remodel experiment-specific confounders, as these have been accounted for in the statistical tests that produced the sets of p-values. Finally, the size of the combined dataset is reduced, potentially reducing the high-dimensionality problem for subsequent integration methods, and related to this, evidence combination is also computationally relatively simple. In sum, such methods could find application in integrative settings which seek to identify functionally relevant signals for prioritisation, where the direction of causality is

perhaps not of immediate importance.

5.4.1 Application of selected integrative methods to ASD

Next, we carried out an integrative analysis of our gene expression and methylation ASD dataset following the suggested analysis flow and a selection of joint analysis methods. Confirmatory analysis was first performed which revealed modest correlation between the measurements on both platforms, and between loci identified as showing significant differences between ASD cases and controls across the entire sample. Following this, two different inferential methods were applied with the aim of uncovering meta-dimensional associations. The first of these was a meta-analytic approach that involved two subsequent p-value combination steps, where evidence was combined firstly at the level of the individual datasets, using the SLK method to account for intraset correlation in the methylation data, and then across datasets, using empirical Brown's method to account for interset correlation (where raw experimental measurements were available from both assays). This produced a combined results list from the separate expression and methylation studies. For the second approach, the nominally significant trait-associated genes from the combined results list were taken forward and a regression-based model selection procedure used to identify functionally relevant signals in the form of likely *cis* methQTLs. The results of model selection where then combined with measures of correlation (previously generated in the confirmatory stage of the analysis), in order to provide a final set of meta-dimensional associations showing evidence for significant association from both the regression based and p-combination based methods, and high levels of correlation between the methylation and expression values.

The final list of nominally significant integrative results revealed a number of interesting and promising candidates. To begin with, *NLGN2* was identified, coding for a neuroligin involved in synaptic cell adhesion. We believe that this could represent a novel ASD risk gene, based on prior evidence of the association of other neuroligin family genes with ASD, such as *NLGN1* [45], *NLGN3* and *NLGN4* [46]. Further, the *NLGN2* gene has also been associated with schizophrenia [47], which has

previously been shown to have a strong genetic overlap with ASD [48, 49, 50]. Another gene of interest, the ubiquitin coding gene *UBE3A*, is a well established ASD susceptibility locus, having been identified as causative for Angelman syndrome, as well as lying in an chromosomal region subject to duplication events frequently associated with idiopathic ASD [51, 52]. This gene is expressed exclusively from the maternal allele, with the paternal allele silenced via epigenetic mechanisms, with disruption to these parent-of-origin specific patterns being causative for Angelman [53]. As there is evidence for both methylation and overall expression of this gene being relevant to ASD, it is therefore extremely encouraging that integrative analysis identified this as a functionally relevant signal, even if this does not constitute a novel finding. Finally, *OXTR* was identified which codes for the oxytocin receptor *OXTR* and is a well established ASD risk gene [54]. Previous studies have also indicated that epigenetic and gene regulatory effects at this locus are associated with ASD [55]. One previous study using cortex tissue samples from 8 ASD cases and 8 matched controls found statistically significant hypermethylation of the *OXTR* gene along with an associated decrease in mRNA levels in samples from a subset of these individuals [55]. Once again, while this does not constitute a novel finding, the fact that the gene emerged here offers further support for that the integrative approach being used here is able to identify genuine meta-dimensional signals.

It is perhaps worth noting that these 3 genes did not appear in the top 50-100 lists from the individual differential expression and differential methylation analyses, and achieved nominal significance in one or other (but not both) of the datasets. The integrative approach used here identified these as significant meta-dimensional associations, with further evidence for functional relevance (in terms of likely gene regulatory effects), which is corroborated by the evidence from other studies. We believe this demonstrates the potential power of such methods, particularly when dealing with genomics data generated by small family-based studies which are likely to be individually underpowered to reveal robust signals.

In terms of potential limitations, our analysis focused on proximal, *cis* regulatory effects. While this simplified the analysis, it means that we are potentially missing important *trans* regulatory effects, as for example, it is known that imprinting

control centers can regulate the expression of genes several megabases away [56]. Investigation of distal relationships between CpG sites and gene expression would require a much larger number of tests to be performed, which must then be accounted for in significance calculation using both regression and p-value combination based methods. This is less of an issue when limiting analysis to *cis* sites, as then only a relatively small number of tests are performed. Any future study intending to examine distal regulation would have to address the issue of multiple testing, perhaps using permutation-based strategy.

Another limitation of this study is the lack of integration with genetic data. There are two main ways in which we might choose to include genetic data to enrich the analysis. Firstly, by utilising genetic data on the same sample of twins, we include another dimension in which to identify associations. We could also then potentially ask further questions about the direction of the effects by using methods such as Mendelian Randomization to tease apart the relationship between disrupted methlyation, gene expression, and ASD. Secondly, we could use public data in order to better place the findings within their functional context in relation to ASD. For example, by examining the intersection between our findings and those from ASD GWAS we could identify putative disorder-associated eQTLs - that is, genes shown by our integrative analysis to have a regulatory signal that also harbour known ASD risk variants.

5.4.2 Future directions

For integrative genomics in general, because such a wide range of different statistical methods and analytical frameworks are available, there needs to be an empirical assessment of the effectiveness of the different approaches for addressing particular classes of problem and types of data. To this end, a sensitivity study for a selection of integrative approaches using a model gene regulatory system involved in disease, with well characterised relationships between genetic variation, gene expression, and epigenetic mechanisms, would be of great value to the wider research community. As for data integration in the context of the ASD MZ twins cohort, for future studies we may look to incorporate further sources of data and

attempt to perform further integrative analyses. For example, exome chip data has recently been generated, which could be used to perform causal inference, using the genetic data to establish the directionality of the expression, methylation and trait interactions.

References

[1] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," *PLoS One*, vol. 6, no. 11, p. e24709, 2011.

[2] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, C. R. Miller, L. Ding, T. Golub, J. P. Mesirov, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer cell*, vol. 17, no. 1, pp. 98–110, 2010.

[3] S. Yoo, S. Takikawa, P. Geraghty, C. Argmann, J. Campbell, L. Lin, T. Huang, Z. Tu, R. Feronjy, A. Spira, *et al.*, "Integrative analysis of dna methylation and gene expression data identifies epas1 as a key regulator of copd," *PLoS Genet*, vol. 11, no. 1, p. e1004898, 2015.

[4] H. Zhong, J. Beaulaurier, P. Y. Lum, C. Molony, X. Yang, D. J. MacNeil, D. T. Weingarth, B. Zhang, D. Greenawalt, R. Dobrin, *et al.*, "Liver and adipose expression associated snps are enriched for association to type 2 diabetes," *PLoS Genet*, vol. 6, no. 5, p. e1000932, 2010.

[5] J. S. Hamid, P. Hu, N. M. Roslin, V. Ling, C. M. Greenwood, and J. Beyene, "Data integration in genetics and genomics: methods and challenges," *Human genomics and proteomics*, vol. 1, no. 1, 2009.

[6] R. D. Hawkins, G. C. Hon, and B. Ren, "Next-generation genomics: an integrative approach," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 476–486, 2010.

[7] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[8] L. Kannan, M. Ramos, A. Re, N. El-Hachem, Z. Safikhani, D. M. Gendoo, S. Davis, D. Gomez-Cabrero, R. Castelo, K. D. Hansen, *et al.*, "Public data and open source tools for multi-assay genomic investigation of disease," *Briefings in bioinformatics*, p. bbv080, 2015.

[9] E. P. Consortium *et al.*, "The encode (encyclopedia of dna elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[10] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, *et al.*, "Gateways to the fantom5 promoter level mammalian expression atlas," *Genome biology*, vol. 16, no. 1, pp. 1–14, 2015.

[11] G. C. Tseng, D. Ghosh, and X. J. Zhou, *Integrating Omics Data*. Cambridge University Press, 2015.

[12] E. Schadt, "Novel integrative genomics strategies to identify genes for complex traits," *Animal genetics*, vol. 37, no. s1, pp. 18–23, 2006.

[13] S. Pineda, P. Gomez-Rubio, A. Picornell, K. Bessonov, M. Márquez, M. Kogevinas, F. Real, K. Van Steen, and N. Malats, "Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases," *Human heredity*, vol. 79, no. 3-4, pp. 124–136, 2015.

[14] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi, "Methods for the integration of multi-omics data: mathematical aspects," *BMC Bioinformatics*, vol. 17, no. 2, p. 167, 2016.

[15] V. N. Kristensen, O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale, "Principles and methods of integrative genomic analyses in cancer," *Nature Reviews Cancer*, vol. 14, no. 5, pp. 299–313, 2014.

[16] Y. Wei, "Integrative analyses of cancer data: A review from a statistical perspective," *Cancer informatics*, vol. 14, no. Suppl 2, p. 173, 2015.

[17] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, *et al.*, "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature genetics*, vol. 37, no. 7, pp. 710–717, 2005.

[18] C. L. Relton and G. D. Smith, "Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease," *International journal of epidemiology*, vol. 41, no. 1, pp. 161–176, 2012.

[19] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *The Review of Economic and Statistics*, pp. 92–107, 1967.

[20] J.-K. Rhee, K. Kim, H. Chae, J. Evans, P. Yan, B.-T. Zhang, J. Gray, P. Spellman, T. H.-M. Huang, K. P. Nephew, *et al.*, "Integrated analysis of genome-wide dna methylation and gene expression profiles in molecular subtypes of breast cancer," *Nucleic acids research*, vol. 41, no. 18, pp. 8464–8474, 2013.

[21] M. Zhao, S. Liu, S. Luo, H. Wu, M. Tang, W. Cheng, Q. Zhang, P. Zhang, X. Yu, Y. Xia, *et al.*, "Dna methylation and mrna and microrna expression of sle cd4+ t cells correlate with disease phenotype," *Journal of autoimmunity*, vol. 54, pp. 127–136, 2014.

[22] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[23] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, p. kxp008, 2009.

[24] I. González, K. Lê Cao, and S. Déjean, "Mixomics: Omics data integration project," *URL http://www. math. univ-toulouse. fr/˜ biostat/mixOmics*, 2011.

[25] C. G. Bell, S. Finer, C. M. Lindgren, G. A. Wilson, V. K. Rakyan, A. E. Teschendorff, P. Akan, E. Stupka, T. A. Down, I. Prokopenko, *et al.*, "Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the fto type 2 diabetes and obesity susceptibility locus," *PloS one*, vol. 5, no. 11, p. e14040, 2010.

[26] O. Gevaert, R. Tibshirani, and S. K. Plevritis, "Pancancer analysis of dna methylation-driven genes using methylmix," *Genome Biol*, vol. 16, no. 1, p. 17, 2015.

[27] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in bioinformatics*, vol. 8, no. 1, pp. 32–44, 2007.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[29] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[30] S. Pineda, F. X. Real, M. Kogevinas, A. Carrato, S. J. Chanock, N. Malats, and K. Van Steen, "Integration analysis of three omics data using penalized regression methods: An application to bladder cancer," *PLoS Genet*, vol. 11, no. 12, p. e1005689, 2015.

[31] A.-L. Boulesteix and K. Strimmer, "Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach," *Theoretical Biology and Medical Modelling*, vol. 2, no. 1, p. 23, 2005.

[32] L. V. Hedges and I. Olkin, *Statistical methods for meta-analysis*. Academic press, 2014.

[33] A. Mishra and S. Macgregor, "Vegas2: Software for more flexible gene-based testing," *Twin Research and Human Genetics*, vol. 18, no. 01, pp. 86–91, 2015.

[34] F. Dudbridge and B. P. Koeleman, "Rank truncated product of p-values, with application to genomewide association scans," *Genetic epidemiology*, vol. 25, no. 4, pp. 360–366, 2003.

[35] K. Yu, Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee, "Pathway analysis by adaptive combination of p-values," *Genetic epidemiology*, vol. 33, no. 8, pp. 700–709, 2009.

[36] Y. Benjamini and R. Heller, "Screening for partial conjunction hypotheses," *Biometrics*, vol. 64, no. 4, pp. 1215–1222, 2008.

[37] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational statistics & data analysis*, vol. 47, no. 3, pp. 467–485, 2004.

[38] S. Won, N. Morris, Q. Lu, and R. C. Elston, "Choosing an optimal method to combine p-values," *Statistics in medicine*, vol. 28, no. 11, pp. 1537–1553, 2009.

[39] J. T. Kost and M. P. McDermott, "Combining dependent p-values," *Statistics & Probability Letters*, vol. 60, no. 2, pp. 183–190, 2002.

[40] W. Poole, D. L. Gibbs, I. Shmulevich, B. Bernard, and T. Knijnenburg, "Combining dependent p-values with an empirical adaptation of brown's method," *bioRxiv*, p. 029637, 2015.

[41] K. J. Kechris, B. Biehs, and T. B. Kornberg, "Generalizing moving averages for tiling arrays using combined p-value statistics," *Statistical applications in genetics and molecular biology*, vol. 9, no. 1, 2010.

[42] B. S. Pedersen, D. A. Schwartz, I. V. Yang, and K. J. Kechris, "Comb-p: software for combining, analyzing, grouping and correcting spatially correlated p-values," *Bioinformatics*, vol. 28, no. 22, pp. 2986–2988, 2012.

[43] C. Wong, E. L. Meaburn, A. Ronald, T. Price, A. Jeffries, L. C. Schalkwyk, R. Plomin, and J. Mill, "Methylomic analysis of monozygotic twins discordant for autism spectrum disorder and related behavioural traits," *Molecular psychiatry*, vol. 19, no. 4, p. 495, 2014.

[44] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[45] J. T. Glessner, K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, *et al.*, "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes," *Nature*, vol. 459, no. 7246, pp. 569–573, 2009.

[46] S. Jamain, H. Quach, C. Betancur, M. Råstam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, *et al.*, "Mutations of the x-linked genes encoding neuroligins nlgn3 and nlgn4 are associated with autism," *Nature genetics*, vol. 34, no. 1, pp. 27–29, 2003.

[47] C. Sun, M.-C. Cheng, R. Qin, D.-L. Liao, T.-T. Chen, F.-J. Koong, G. Chen, and C.-H. Chen, "Identification and functional characterization of rare mutations of the neuroligin-2 gene (nlgn2) associated with schizophrenia," *Human molecular genetics*, vol. 20, no. 15, pp. 3042–3051, 2011.

[48] L. S. Carroll and M. J. Owen, "Genetic overlap between autism, schizophrenia and bipolar disorder," *Genome medicine*, vol. 1, no. 10, p. 1, 2009.

[49] B. Crespi, P. Stead, and M. Elliot, "Comparative genomics of autism and schizophrenia," *Proceedings of the National Academy of Sciences*, vol. 107, no. suppl 1, pp. 1736–1741, 2010.

[50] C.-D. G. of the Psychiatric Genomics Consortium *et al.*, "Genetic relationship between five psychiatric disorders estimated from genome-wide snps," *Nature genetics*, vol. 45, no. 9, pp. 984–994, 2013.

[51] P. F. Bolton, N. Dennis, C. Browne, N. Thomas, M. Veltman, R. Thompson, and P. Jacobs, "The phenotypic manifestations of interstitial duplications of proximal 15q with special reference to the autistic spectrum disorders," *American journal of medical genetics*, vol. 105, no. 8, pp. 675–685, 2001.

[52] S. E. Roberts, N. R. Dennis, C. E. Browne, L. Willatt, G. C. Woods, I. Cross, P. A. Jacobs, and S. N. Thomas, "Characterisation of interstitial duplications

and triplications of chromosome 15q11–q13," *Human genetics*, vol. 110, no. 3, pp. 227–234, 2002.

[53] Y.-h. Jiang, T. Sahoo, R. C. Michaelis, D. Bercovich, J. Bressler, C. D. Kashork, Q. Liu, L. G. Shaffer, R. J. Schroer, D. W. Stockton, *et al.*, "A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for ube3a," *American Journal of Medical Genetics Part A*, vol. 131, no. 1, pp. 1–10, 2004.

[54] S. Jacob, C. W. Brune, C. Carter, B. L. Leventhal, C. Lord, and E. H. Cook, "Association of the oxytocin receptor gene (oxtr) in caucasian children and adolescents with autism," *Neuroscience letters*, vol. 417, no. 1, pp. 6–9, 2007.

[55] S. G. Gregory, J. J. Connelly, A. J. Towers, J. Johnson, D. Biscocho, C. A. Markunas, C. Lintas, R. K. Abramson, H. H. Wright, P. Ellis, *et al.*, "Genomic and epigenetic evidence for oxytocin receptor deficiency in autism," *BMC medicine*, vol. 7, no. 1, p. 1, 2009.

[56] A. Lewis and W. Reik, "How imprinting centres work," *Cytogenetic and genome research*, vol. 113, no. 1-4, pp. 81–89, 2006.

6 Discussion

6.1 Gene expression signatures in ASD

In what we believe represents the largest study of its kind performed to date, we generated gene expression profiles from whole-blood samples taken from a cohort of ASD concordant and discordant MZ twins using high resolution RNA-seq. By comparing expression profiles between ASD affected and unaffected individuals across the entire sample we were able to identify differences that could be genetically or non-genetically mediated, and by comparing expression within discordant pairs, we were able to identify changes in gene expression likely to reflect non shared environmental factors. Overall, we established convergent evidence for a number of genes with potential relevance to ASD: *DEPDC1B*, *IGHG4, IGHG3, IGHV3-66*, *HSPA8P14, HSPA13*, *SLC15A2*. Pathway analysis showed enrichment for genes involved in transcriptional control, immunity and *PI3K/AKT* signalling. The immune-related genes were of particular interest, due to the fact that they were also present in the discordant MZ analysis, and showed a consistent direction of differential expression, being up-regulated in cases compared to unaffected twins. Although there are many limitations to this study (as covered in the relevant Discussions in Chapters 2 and 3) there are reasons to be confident that these may include genuine associations. To begin with, the overwhelming majority of previous gene expression studies into ASD have also implicated immune system and inflammatory pathways (see Chapter 1 - Table 1). In addition, there is also wealth of evidence for altered immune response in ASD that comes from a number of epidemiological, serological, imaging, and postmortem studies (see [1, 2]

for reviews).

6.1.1 Future directions

While we hope that this work goes some way towards demonstrating the exciting potential of functional genomics investigations utilizing the MZ twins paradigm in order to tease apart genes from environment and identify molecular biomarkers in ASD, we acknowledge that further work needs to be undertaken to really establish the veracity and translational relevance of the findings.

We have previously discussed the various limitations and shortcomings of the study, but just to re-iterate, one major issue is the use of whole-blood as a proxy for brain which remains controversial. As was suggested in the Discussion section of Chapter 4, one way in which we might address this is by comparing our results to those from the GTEx project, to determine which of the identified signals is also highly expressed in the brain, and hence likely to be disorder-relevant. However, this would primarily serve as a way of annotating the results for follow up but would not directly test whether differences in expression are also present in the brain and are also associated with ASD, which is really the key question. In order to address this, ideally we would attempt to replicate our findings in postmortem brain tissue resource, but a major limitation is that samples are unlikely to be derived from MZ concordant and discordant twins, and so we lose an important aspect of our design here, which is the genetic matching of MZ twins to allow for investigation of expression differences mediated by non-shared environment. Such a study would then be primarily useful for replicating those differences which are likely to be genetically driven. Alternatively, we could look at the possibility of profiling other tissues in our cohort, in an attempt to get closer to the primary affected tissue. For example, scalp hair follicles are non-invasive to collect and are derived from the ectoderm, the same germ layer as brain tissue. We would perhaps be more confident of detecting gene expression (and indeed epigenetic) changes that have arisen early in development and are hence more likely to be ASD relevant. Indeed, study by Maekawa et al. investigated the utility of scalp hair follicles in schizophrenia and ASD and identified a putative novel biomarker in schizophrenia, as well as finding

that in ASD that *CNTNAP2*, a strong ASD candidate gene, was significantly down-regulated in follicles [3].

Another related issue which was also discussed was the fact that the blood samples were collected from adolescent subjects, and hence we are unable to say whether the identified expression signatures in blood are causal for ASD. Since the pathogenesis of ASD occurs during early development, identified signals could be a response to the initial perturbation - this would seem particularly plausible in the case of the immune signal identified. In which case, it is important to keep in mind that with a sample such as this, we are in some ways looking at the shadows cast by an earlier developmental disruption, making it difficult to draw inferences about causal pathways. Perhaps the primary goal of any gene expression profiling using peripheral tissue should be the identification of biomarkers, with our cohort this would be for diagnostic purposes and to help distinguish environmentally sensitive markers, but with a study cohort where samples were collected at an earlier stage this could also potentially yield predictive and prognostic markers. Indeed, the accumulated evidence from other gene expression profiling studies, which consistently identify immune and transcriptional control pathways, suggests the existence of a blood-based expression signature of ASD. While we are hopeful that refinement of the methods used here can lead to the identification of a robust signature for use in a clinical setting, from the outset this would seem to be a less than ideal way of investigating molecular mechanisms, for the reasons stated. Studies aiming to contribute to our mechanistic understanding of ASD could then perhaps look to tissues other than blood (as discussed above) and ideally from prospective birth cohorts. There could also be utility in utilizing iPSCs (induced pluripotent stem cells), again to try and recreate developmental conditions as closely as possible using any available, non-invasive peripheral samples such as hair follicle.

Finally, on the issue of teasing apart genetic from environmentally mediated gene expression alterations in ASD, it has previously been suggested that the repeated identification of immune and transcriptional control related pathways suggests that these might be sensitive to environmental influence [4, 5]. Our findings here would seem to offer further support this hypothesis. Since extensive phenotypic measures are available for our cohort, and genetic data will become available in

the near future, an interesting follow up study would be perform causal analysis in an attempt to link disrupted expression and methylation within discordant pairs to any early-life environmental exposures. It might then be possible to link for example, early viral infection with up-regulation of immune genes, and a pro-inflammatory state interfering with development at a critical period and leading to the development of ASD.

6.2 Significance thresholds and sample sizes required for EWAS

We also addressed a methodological issue relevant to epigenome-wide association studies by deriving an estimate for significance for a single site to be declared as differentially methylated using the 450K array: $\alpha = 2.5 \times 10^{-7}$, which was then extrapolated to obtain a genome-wide estimate: $\alpha = 4.5 \times 10^{-8}$. It was initially anticipated that this estimate would be useful in our study for ensuring the robustness of the results from the methylation dataset being taken forward for integration with the gene expression dataset. However, when it was later realized that this significance estimate is likely to be platform specific and the results of extrapolation perhaps not reliable, this threshold was not applied. The work does however remain relevant to the ASD study, as we then also used the estimated thresholds to calculate the sample sizes required for a methylomic profiling experiment. Here, it was found that to detect differences of 10% in methylation levels with 80% power at the genome-wide significance level, a twin-based design would require ~ 60 pairs of twins. Given the functional link between methylation and expression, this could perhaps suggest that larger sample sizes might also be required for gene expression studies, especially given that the log fold changes within twin pairs that we observed tended to be of a similar magnitude. It might also suggest that our gene expression study with 5 discordant pairs, was in hindsight underpowered to detect robust, highly significant differences in gene expression - although we later demonstrated that it might be possible to compensate for this by using integrative methods.

6.3 Integration of gene expression and methylation data

The gene expression profiles generated here were then integrated with methylation profiles on the same MZ twin cohort. Key findings from integration were the identification of *NLGN2*, *UBE3A*, *OXTR* as showing combined evidence for dysregulation in ASD. These failed to reach significance in the individual datasets, yet by applying integrative methods they achieved nominal, combined significance. That *UBE3A* and *OXTR* were identified is particularly encouraging, as previous methylomic and expression studies have identified these genes as being differentially methylated or differentially expressed in ASD [6, 7]. So while these might not represent novel findings, they do at least suggest that the integrative approach developed here has had some degree of succes in identifying ASD-associated genes that are likely to display epigenetic and gene expression differences. This increases confidence in the findings, and suggests that other potentially relevant genes might also have been identified. Future studies could prioritise some of these other genes for follow up in order to identify novel candidates. Here, we suggested that *NLGN2* in particular would be a good candidate, as mutations in genes coding for a number of other neuroligins have previously been associated with ASD [8, 9].

6.3.1 Future directions

We have generated a rich, multi-dimensional ASD dataset which could be used for a number of interesting future research projects. One of the biggest challenges in ASD research is in translating findings in order to improve the accuracy of diagnosis, and make better predictions about response to behavioural and potentially pharmacological interventions. Here, we found evidence for an immune signature in whole blood, which could have potential as a diagnostic biomarker. Since such an immunologically focused, blood-based, gene expression test has in fact already been developed [10], in order to introduce a novel aspect, we might wish to leverage our additional data sources to search for a multidimensional methylation and

expression signature and see whether this improves diagnostic power.

Another issue that studies into ASD must contend with is phenotypic heterogeneity, which is believed to be related to the underlying genetic heterogeneity [11]. It has been suggested this could be one of the main reasons behind the lack of robust association findings from ASD GWAS. A number of studies have suggested using phenotypic or biological endophenotypes to reduce heterogeneity and improve power, for example using trait severity [12], or subgroups based on identified genetic risks [11]. Future studies may wish to explore the idea of using a multidimensional classifier in order to identify ASD endophenotypes across the different molecular layers. This could potentially allow cases to be stratified according to the primary pathway perturbation - for example: synaptic organisation, synaptic function, immune response, transcriptional control. Finally, integrative approaches might also be used to better dissect etiology, by using the available data to obtain multidimensional risk scores based on the burden of previously identified risks across genetic, epigenetic, and gene expression datasets.

6.4 Conclusions

Molecular genetics aims to progress our understanding of ASD from the bottom-up, by identifying the genes and pathways likely to be involved and then using this to build up a mechanistic understanding. Two key challenges for molecular ASD research are i) to develop a coherent model for population risk, ii) devise a unified molecular model for pathology that explains the disparate genomics findings to date and incorporates commonly observed findings and popular pathological explanations such as synaptic dysfunction, altered neural circuitry, inhibitory/excitatory imbalance, transcriptional dysregulation, and inflammation. In this thesis, we contributed to the second of these challenges, in attempting to find pathways upon which risk factors converge that are genetically or potentially environmentally sensitive, by identifying disorder-associated signals in gene expression and DNA methylation datasets from MZ twins both concordant and discordant for ASD. There is increasing evidence that establishment and maintenance of methylation marks

during development and coordination of gene expression is important in the context of normal brain development. This, taken together with the accumulated evidence from genetic and functional studies showing the involvement of genes and pathways involved in transcriptional regulation in ASD, and that early environmental exposures during development can leave an imprint on the epigenome, suggests a possible multi-etiological risk pathway for ASD involving genes, epigenetic regulation of gene expression, and early environmental insults. Our results lend support to the hypothesis that disruption of activity dependent transcriptional/translational could be a key component of the molecular chain of events leading to ASD, given the importance of gene expression flexibility during development for forming and pruning synaptic connections [13]. Altered gene expression dynamics could potentially lead to disruption of connectivity in higher centers of the brain important for mediating social behaviours (e.g. frontal-parietal, frontal temporal, frontal-striatal circuits) [14]. The possibility that regulatory disruption could serve as an integrator of early genetic and environmental risk factors justifies further investigation into disrupted epigenetic and gene expression in ASD. In order to explore such hypotheses more fully, we should look to integrative genomics approaches in order to consider the different molecular layers in combination, helping us to gain a more systems-level perspective on autism.

References

[1] D. L. Vargas, C. Nascimbene, C. Krishnan, A. W. Zimmerman, and C. A. Pardo, "Neuroglial activation and neuroinflammation in the brain of patients with autism," *Annals of neurology*, vol. 57, no. 1, pp. 67–81, 2005.

[2] P. Ashwood, S. Wills, and J. Van de Water, "The immune response in autism: a new frontier for autism research," *Journal of leukocyte biology*, vol. 80, no. 1, pp. 1–15, 2006.

[3] M. Maekawa, K. Yamada, M. Toyoshima, T. Ohnishi, Y. Iwayama, C. Shimamoto, T. Toyota, Y. Nozaki, S. Balan, H. Matsuzaki, *et al.*, "Utility of scalp hair follicles as a novel source of biomarker genes for psychiatric illnesses," *Biological psychiatry*, vol. 78, no. 2, pp. 116–125, 2015.

[4] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[5] V. W. Hu, "From genes to environment: Using integrative genomics to build a "systems-level" understanding of autism spectrum disorders," *Child development*, vol. 84, no. 1, pp. 89–103, 2013.

[6] Y.-h. Jiang, T. Sahoo, R. C. Michaelis, D. Bercovich, J. Bressler, C. D. Kashork, Q. Liu, L. G. Shaffer, R. J. Schroer, D. W. Stockton, *et al.*, "A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role

for ube3a," *American Journal of Medical Genetics Part A*, vol. 131, no. 1, pp. 1–10, 2004.

[7] S. G. Gregory, J. J. Connelly, A. J. Towers, J. Johnson, D. Biscocho, C. A. Markunas, C. Lintas, R. K. Abramson, H. H. Wright, P. Ellis, *et al.*, "Genomic and epigenetic evidence for oxytocin receptor deficiency in autism," *BMC medicine*, vol. 7, no. 1, p. 1, 2009.

[8] S. Jamain, H. Quach, C. Betancur, M. Råstam, C. Colineaux, I. C. Gillberg, H. Soderstrom, B. Giros, M. Leboyer, C. Gillberg, *et al.*, "Mutations of the x-linked genes encoding neuroligins nlgn3 and nlgn4 are associated with autism," *Nature genetics*, vol. 34, no. 1, pp. 27–29, 2003.

[9] J. T. Glessner, K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, *et al.*, "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes," *Nature*, vol. 459, no. 7246, pp. 569–573, 2009.

[10] S. J. Glatt, M. T. Tsuang, M. Winn, S. D. Chandler, M. Collins, L. Lopez, M. Weinfeld, C. Carter, N. Schork, K. Pierce, *et al.*, "Blood-based gene expression signatures of infants and toddlers with autism," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, no. 9, pp. 934–944, 2012.

[11] S. S. Jeste and D. H. Geschwind, "Disentangling the heterogeneity of autism spectrum disorder through genetic findings," *Nature Reviews Neurology*, vol. 10, no. 2, pp. 74–81, 2014.

[12] O. Veatch, J. Veenstra-VanderWeele, M. Potter, M. A. Pericak-Vance, and J. Haines, "Genetically meaningful phenotypic subgroups in autism spectrum disorders," *Genes, Brain and Behavior*, vol. 13, no. 3, pp. 276–285, 2014.

[13] E. Yeh and L. A. Weiss, "If genetic variation could talk: What genomic data may teach us about the importance of gene expression regulation in the genetics of autism," *Molecular and Cellular Probes*, vol. 30, no. 6, pp. 346–356, 2016.

[14] B. S. Abrahams and D. H. Geschwind, "Connecting genes to brain in the autism spectrum disorders," *Archives of neurology*, vol. 67, no. 4, pp. 395–399, 2010.

Appendix

A Chapter 2

```r
twoCM <- function(x,c){

        # R function for fitting a two component mixture model to array intensity data
        # for filtering out probes measuring background noise (representing smaller component)
        # x is an oligo expression set object,
        # c is the confidence level - probes that can be assigned to smaller
        # noise distribution with this level of confidence

        require('mixtools')

        #collapse expression data into a single vector
        exprs <- vector()
        for (i in 1:ncol(exprs(x))) {
                exprs <- c(exprs,exprs(x)[,i])
        }

        #run normalmix EM procedure
        exprs_mix <- normalmixEM(exprs, k=2,epsilon = 1e-08, maxit = 1000, verb=TRUE,
                                                fast=T, arbmean= T,arbvar = T)

        #get posteriors
        post_prob <- as.matrix(exprs_mix$posterior)

        #keep only posteriors for smaller component (defined as one with smaller mean)
        #and reconstitute into matrix
        m <- dim(exprs(x))[1]
        n <- dim(exprs(x))[2]
        mix_matrix <- matrix(nrow=m, ncol=n)
        for (i in (1:n)) {
                mix_matrix[,i] <- post_prob[(((i*m)-m)+1):((i*m)),which.min(exprs_mix$mu)]
        }

        #calculate mean posteriors for each gene
        post_means <- apply(mix_matrix, 1, function(x) mean(x))
        post_means <- as.matrix(post_means)

        #omit genes that can be assigned to smaller distribution with confidence > n%
        omit_probes <- which(post_means > c)
        return(x[-omit_probes,])
}
```

Figure A1. *R* function to fit a two component mixture model to microarray data

B Chapter 3

Experimental Design - Lane assignments

P448

Key
Group 1
Group 2
Group 6

LANE 1	LANE 2	LANE 3	LANE 4	LANE 5	LANE 6	LANE 7	LANE 8
TG98582	TG78042	TG74382	TG78041	TG36391	TG74381	TG36392	TG29291
TG104611	TG236001	TG200022	TG123521	TG139622	TG48482	TG200021	TG104612
TG17181	TG66562	TG66561	TG129672	TG126552	TG15321	TG15081	TG20481
TG46142	TG126551	TG19901	TG139621	TG17571	TG98581	TG49452	TG49451
TG129671	TG20482	TG15082	TG236002	TG113631	TG15322	TG48481	TG113632

Table B1. Flow cell lane assignments for the RNA-seq experiment

Pre-alignment QC

					BioAnalyzer		FastQC (pre-trim)		SAM tools			Picard tools			
	Sample	Gender	Group	Lane	Library size(bp)	RIN	Read length (bp)	Total Reads	GC %	Post-trimmed Reads	Properly Paired	Median insert size (bp)	Insert size SD	Duplication rate	Mate inner distance
P448												Tissue type: Whole-blood, Library kit: TruSeq Total RNA, Sequencing Platform: Illumina Hi-Seq 2000 paired end			
	TG98582	M	1	1	NA	8.2	101	49,878,602	46	54,313,294	27,249,112	175	94	0.19	-27
	TG104611	M	2	1	NA	4.13	101	66,983,208	49	67,591,455	35,603,340	169	86	0.21	-33
	TG17181	F	6	1	NA	7.7	101	46,538,044	47	52,910,477	27,110,594	181	100	0.23	-21
	TG46142	M	1	1	NA	8.4	101	46,247,316	47	51,715,190	26,373,198	173	88	0.23	-29
	TG129671	M	6	1	NA	NA	101	51,502,760	47	36,282,748	18,842,922	175	95	0.24	-27
	TG78042-REP-L002	M	1	2	NA	8.4	101	52,929,574	46	55,935,851	28,958,120	172	90	0.16	-30
	TG236001	F	2	2	NA	6.9	101	75,193,360	47	82,530,717	42,309,884	170	86	0.19	-32
	TG66562	F	6	2	NA	8.1	101	59,129,470	47	69,599,533	35,466,788	175	88	0.24	-27
	TG126551	F	6	2	NA	NA	101	54,719,570	46	64,715,175	33,037,216	183	103	0.25	-19
	TG20482	M	6	2	NA	7.8	101	58,716,442	46	38,444,405	18,462,628	181	95	0.14	-21
	TG74382	M	1	3	NA	8.7	101	61,614,838	47	25,204,814	12,171,012	183	97	0.17	-19
	TG200022	F	2	3	NA	7.6	101	66,007,688	46	70,018,926	33,890,386	181	96	0.20	-21
	TG66561	F	6	3	NA	8.1	101	66,007,688	46	85,789,089	44,845,578	169	84	0.26	-33
	TG19901	F	6	3	NA	8.6	101	64,217,692	48	46,730,873	23,813,958	177	96	0.26	-25
	TG15082	F	6	3	NA	8.8	101	70,792,894	47	80,178,916	40,783,982	175	94	0.26	-27
	TG78041	M	1	4	NA	8	101	43,675,538	47	48,002,965	24,012,374	175	91	0.21	-27
	TG123521	F	2	4	NA	7.6	101	55,392,432	46	60,269,977	29,184,726	185	100	0.18	-17
	TG129672	M	6	4	NA	NA	101	55,094,368	48	48,243,438	25,197,434	174	96	0.26	-28
	TG139621	M	2	4	NA	8.8	101	45,365,368	47	30,505,529	15,101,578	179	96	0.14	-23
	TG236002	F	2	4	NA	8.2	101	47,034,014	48	28,318,691	14,353,394	177	92	0.22	-25
	TG36391	M	1	5	NA	6.8	101	54,023,076	47	57,574,785	28,058,412	173	86	0.14	-29
	TG139622	M	2	5	NA	9	101	49,249,946	46	52,166,754	24,331,170	189	106	0.21	-13
	TG126552	F	6	5	NA	NA	101	55,489,936	47	62,526,415	32,623,702	174	92	0.26	-28
	TG17571	M	6	5	NA	8.8	101	56,069,690	47	35,507,100	17,634,376	176	92	0.16	-26
	TG113631	F	6	5	NA	8.6	101	62,739,468	47	73,074,029	35,170,354	182	96	0.26	-20
	TG74381	M	1	6	NA	8.6	101	61,241,626	47	22,335,499	11,464,580	176	91	0.19	-26
	TG48482	M	2	6	NA	7.9	101	64,396,234	48	61,939,552	30,564,012	178	95	0.24	-24
	TG15321	M	6	6	NA	8.4	101	59,004,344	48	48,352,495	25,011,632	172	93	0.21	-30
	TG98581	M	1	6	NA	8.3	101	54,051,952	46	59,165,002	29,118,480	182	99	0.21	-20
	TG15322	M	6	6	NA	8.5	101	78,700,222	46	38,649,453	18,882,186	176	98	0.17	-26
	TG78042-REP-L007	M	1	7	NA	8.4	101	38,884,070	46	40,979,058	21,274,970	172	91	0.15	-30
	TG200021	F	2	7	NA	8.3	101	47,613,398	47	53,937,420	27,940,906	174	91	0.25	-28
	TG15081	F	6	7	NA	8.2	101	40,482,952	49	27,358,454	14,080,918	174	94	0.22	-28
	TG49452	F	6	7	NA	7.9	101	39,669,620	47	45,114,982	22,744,352	178	94	0.22	-24
	TG48481	M	2	7	NA	7.9	101	43,282,416	48	50,351,655	25,110,528	187	108	0.27	-15
	TG29291	M	1	8	NA	7.2	101	41,892,950	47	34,947,388	18,052,380	170	87	0.18	-32
	TG104612	M	2	8	NA	8.7	101	45,089,612	47	49,490,056	26,081,388	174	94	0.24	-28
	TG20481	M	6	8	NA	8.5	101	49,114,472	47	27,464,415	14,004,048	174	95	0.17	-28
	TG49451	F	6	8	NA	8.2	101	38,306,698	49	34,533,973	18,140,114	177	92	0.27	-25
	TG113632	F	6	8	NA	8.8	101	40,364,354	48	46,891,155	24,402,376	173	94	0.25	-29

Table B2. Pre-alignment QC metrics for the RNA-seq experiment

Post-alignment QC

		Tophat		RNASeQC														
				General mapping metrics					Transcript- associated reads			Detection rate		Strand specificity		Mean per base coverage		
	Sample	Aligned pairs	Mapping rate	Mapped reads	Uniquely mapped rate	rRNA rate	Mapped pairs	Chimeric pairs	Exonic rate	Intronic rate	Intergenic rate	Transcripts detected	Genes detected	End 1 Sense rate	End 2 Sense rate	Bottom 1000 trans	Middle 1000 trans	Top 1000 trans
P448																		
	TG98582	21,952,196	0.81	45,111,859	0.83	NA	21,945,422	990,208	0.43	0.51	0.06	129,169	24,871	0.04	0.96	6.66	19.56	664.24
	TG104611	27,256,332	0.81	56,089,720	0.81	NA	27,244,722	961,973	0.45	0.50	0.05	132,871	25,593	0.03	0.97	8.65	25.50	932.57
	TG17181	20,455,305	0.77	41,955,146	0.79	NA	20,399,968	953,256	0.49	0.45	0.06	126,175	23,697	0.04	0.96	6.93	20.13	899.12
	TG46142	20,660,140	0.80	42,358,182	0.79	NA	20,655,130	1,062,753	0.45	0.50	0.05	126,109	23,730	0.03	0.97	6.12	17.70	681.71
	TG129671	13,535,778	0.75	27,729,901	0.78	NA	13,523,807	600,967	0.50	0.46	0.05	120,002	21,749	0.03	0.97	4.42	12.69	559.56
	TG78042	22,963,577	0.82	47,205,721	0.86	NA	22,952,125	940,060	0.43	0.51	0.06	129,539	24,746	0.04	0.96	7.49	22.98	831.98
	TG236001	32,452,198	0.79	66,751,817	0.83	NA	32,435,507	1,264,876	0.42	0.51	0.07	134,540	27,121	0.05	0.95	9.76	27.85	1050.95
	TG66562	26,560,779	0.76	54,437,633	0.78	NA	26,555,210	1,345,115	0.46	0.49	0.05	127,054	24,172	0.03	0.97	7.74	21.60	789.71
	TG126551	24,577,762	0.76	50,309,706	0.77	NA	24,568,037	964,360	0.52	0.43	0.05	128,231	24,189	0.03	0.97	8.58	24.66	1013.09
	TG20482	15,727,038	0.82	32,239,667	0.88	NA	15,723,282	820,312	0.43	0.53	0.05	120,869	22,073	0.03	0.97	4.95	14.83	583.03
	TG74382	10,100,033	0.80	20,744,894	0.85	NA	10,097,335	531,903	0.45	0.50	0.05	115,095	20,453	0.03	0.97	3.22	9.64	376.07
	TG200022	27,245,890	0.78	56,012,753	0.83	NA	27,227,507	1,127,163	0.44	0.51	0.05	129,577	24,790	0.03	0.97	8.31	24.30	945.17
	TG66561	32,449,078	0.76	66,589,171	0.75	NA	32,438,036	1,642,945	0.48	0.48	0.05	130,346	25,093	0.03	0.97	9.65	26.91	969.87
	TG19901	17,520,012	0.75	35,934,807	0.76	NA	17,513,795	828,937	0.51	0.45	0.04	123,612	22,721	0.03	0.97	5.76	16.44	636.76
	TG15082	31,017,290	0.77	63,715,361	0.76	NA	31,009,689	1,382,179	0.49	0.47	0.05	129,363	24,873	0.04	0.96	9.72	26.64	1046.22
	TG78041	19,154,673	0.80	39,357,468	0.81	NA	19,147,327	1,049,177	0.47	0.48	0.05	126,406	23,544	0.03	0.97	6.17	17.93	679.69
	TG123521	24,493,115	0.81	50,290,725	0.84	NA	24,486,366	1,170,289	0.49	0.46	0.05	127,571	23,774	0.03	0.97	8.89	26.33	1036.50
	TG129672	17,883,910	0.74	36,617,326	0.76	NA	17,875,914	673,991	0.53	0.43	0.05	123,540	22,787	0.03	0.97	6.08	16.89	755.91
	TG139621	12,729,291	0.83	26,172,091	0.88	NA	12,725,463	554,918	0.46	0.49	0.05	120,886	21,757	0.03	0.97	4.56	14.27	576.83
	TG236002	10,719,330	0.76	21,958,673	0.80	NA	10,714,290	542,476	0.47	0.48	0.05	115,397	20,559	0.03	0.97	3.36	9.39	388.58
	TG36391	23,668,951	0.82	48,686,185	0.89	NA	23,663,079	1,163,737	0.37	0.57	0.05	130,005	24,955	0.03	0.97	6.38	18.85	595.41
	TG139622	21,479,863	0.82	44,131,033	0.82	NA	21,474,115	827,539	0.47	0.48	0.05	127,203	23,863	0.04	0.96	7.17	21.72	797.54
	TG126552	23,963,315	0.77	49,063,556	0.76	NA	23,948,540	823,121	0.50	0.45	0.05	128,811	24,380	0.03	0.97	7.95	22.98	925.20
	TG17571	14,348,758	0.81	29,416,998	0.86	NA	14,345,323	616,356	0.42	0.53	0.05	123,058	22,613	0.03	0.97	4.25	12.57	443.26
	TG113631	27,948,074	0.76	57,294,128	0.76	NA	27,925,833	1,394,006	0.48	0.47	0.05	129,423	24,676	0.03	0.97	8.69	24.41	913.41
	TG74381	8,658,985	0.78	17,763,438	0.83	NA	8,653,772	405,950	0.48	0.47	0.05	112,167	19,661	0.03	0.97	2.88	8.36	350.65
	TG48482	24,032,595	0.78	50,034,010	0.79	NA	24,021,264	1,172,714	0.46	0.47	0.07	127,279	23,880	0.05	0.95	7.82	22.77	955.80
	TG15321	18,240,204	0.75	37,394,083	0.81	NA	18,219,296	718,400	0.50	0.46	0.04	126,142	23,376	0.03	0.97	6.29	17.68	679.98
	TG98581	23,586,084	0.80	48,420,208	0.81	NA	23,577,863	1,079,363	0.50	0.45	0.05	128,620	24,244	0.03	0.97	8.30	25.18	973.50
	TG15322	15,448,460	0.80	31,708,674	0.86	NA	15,441,240	717,170	0.46	0.50	0.05	123,261	22,625	0.03	0.97	5.12	15.28	583.24
	TG78042-REP-L007	16,919,179	0.83	34,741,525	0.87	NA	16,914,326	700,831	0.43	0.51	0.06	125,586	23,282	0.04	0.96	5.59	17.14	657.05
	TG200021	21,008,837	0.78	43,059,799	0.77	NA	21,004,072	1,008,272	0.50	0.45	0.05	125,670	23,476	0.04	0.96	6.80	19.32	778.91
	TG15081	10,561,615	0.77	21,733,410	0.80	NA	10,558,615	480,932	0.49	0.46	0.06	116,325	20,604	0.04	0.96	3.65	10.28	526.62
	TG49452	17,719,234	0.79	36,254,767	0.80	NA	17,715,144	797,486	0.48	0.48	0.05	123,674	22,823	0.03	0.97	5.54	15.52	651.67
	TG48481	18,938,144	0.75	38,789,025	0.75	NA	18,928,949	852,757	0.55	0.40	0.04	123,779	22,603	0.03	0.97	6.96	19.88	852.43
	TG29291	13,939,608	0.80	28,563,057	0.84	NA	13,936,309	534,570	0.43	0.52	0.05	122,045	22,300	0.03	0.97	4.25	12.28	498.50
	TG104612	19,491,305	0.79	39,993,761	0.78	NA	19,485,725	970,512	0.54	0.41	0.05	127,071	23,469	0.03	0.97	7.66	22.83	1028.47
	TG20481	10,907,673	0.79	22,310,813	0.85	NA	10,905,182	393,087	0.46	0.50	0.05	115,962	20,650	0.04	0.97	3.49	10.36	415.69
	TG49451	12,792,062	0.74	26,181,761	0.75	NA	12,788,294	634,562	0.52	0.44	0.04	118,333	21,132	0.03	0.97	4.21	11.17	516.25
	TG113632	17,844,635	0.76	36,554,530	0.76	NA	17,833,505	782,891	0.51	0.44	0.05	122,916	22,634	0.04	0.96	5.98	16.88	708.86

Tissue type: Whole blood, Library kit: TruSeq Total RNA, Sequencing Platform: Illumina HI-Seq 2000 paired end

Table B3. Post-alignment QC metrics for the RNA-seq experiment

Quantification metrics

Sample	Gender	Group	Lane	Total read pairs	% Assigned	Assigned	Unassigned_Ambiguity	Unassigned_MultiMapping	Unassigned_NoFeatures	Un_Unmapped	Un_MappingQuality	Un_FragmentLength	Un_Chimera	Un_Secondary	Un_Nonjunction	Un_Duplicate
TG98582	M	1	1	27,885,131	30.42	8483627	263900	5188667	12275050	0	0	0	1673887	0	0	0
TG104611	M	2	1	34,726,440	31.49	10936655	258053	6908990	14839809	0	0	0	1782933	0	0	0
TG17181	F	6	1	27,087,504	31.39	8502642	174640	6534529	10086755	0	0	0	1788938	0	0	0
TG46142	M	1	1	26,487,759	29.49	7812333	178601	5488708	11119687	0	0	0	1888430	0	0	0
TG129671	M	6	1	18,555,933	29.73	5516109	130571	5208871	6570580	0	0	0	1129802	0	0	0
TG78042-REP-L002	M	1	2	28,716,605	31.64	9085146	216593	4959553	12897096	0	0	0	1558217	0	0	0
TG236001	F	2	2	42,365,827	28.09	11901919	279783	9285755	18521980	0	0	0	2376390	0	0	0
TG66562	F	6	2	35,626,619	27.76	9890272	244793	9056743	13958244	0	0	0	2476567	0	0	0
TG126551	F	6	2	33,079,412	32.21	10655516	286766	8745843	11336909	0	0	0	2054378	0	0	0
TG20482	M	6	2	19,687,806	30.63	6029807	128095	3476910	8775271	0	0	0	1277723	0	0	0
TG74382	M	1	3	12,933,524	30.79	3982687	87155	2550755	5400178	0	0	0	912749	0	0	0
TG200022	F	2	3	35,918,375	28.80	10346251	233055	7995368	15105441	0	0	0	2238260	0	0	0
TG66561	F	6	3	43,977,425	28.39	12486471	298912	11693515	16513016	0	0	0	2985511	0	0	0
TG19901	F	6	3	23,939,732	29.64	7095952	143745	6655845	8391337	0	0	0	1652853	0	0	0
TG15082	F	6	3	41,128,005	30.66	12610416	252111	10060808	15572842	0	0	0	2631828	0	0	0
TG78041	M	1	4	24,635,925	31.31	7714509	181215	5049522	9919884	0	0	0	1770795	0	0	0
TG123521	F	2	4	30,898,665	35.27	10898004	239395	5700054	12122461	0	0	0	1938751	0	0	0
TG129672	M	6	4	24,663,640	30.87	7613898	167982	7282623	8170087	0	0	0	1429050	0	0	0
TG139621	M	2	4	15,659,443	34.45	5394501	118971	2451250	6812115	0	0	0	882606	0	0	0
TG236002	F	2	4	14,485,972	28.08	4067010	81970	3855341	5482933	0	0	0	998718	0	0	0
TG36391	M	1	5	29,577,690	26.72	7903371	178488	4842827	14659909	0	0	0	1993095	0	0	0
TG139622	M	2	5	26,766,868	34.55	9247696	191219	4474314	11241208	0	0	0	1612431	0	0	0
TG126552	F	6	5	31,965,976	32.20	10291765	246839	8129431	11525726	0	0	0	1772215	0	0	0
TG17571	M	6	5	18,184,989	28.90	5255336	120125	3508756	8195230	0	0	0	1105542	0	0	0
TG113631	F	6	5	37,401,457	30.31	11337856	264082	9236701	13962682	0	0	0	2600136	0	0	0
TG74381	M	1	6	11,445,557	30.62	3504379	82971	2747279	4375774	0	0	0	735154	0	0	0
TG48482	M	2	6	32,096,674	29.66	9520833	193845	6994254	13416755	0	0	0	1970987	0	0	0
TG15321	M	6	6	24,747,651	30.82	7627412	192827	6605054	8889592	0	0	0	1432766	0	0	0
TG98581	M	1	6	30,331,007	34.14	10356458	227732	6209643	11546047	0	0	0	1991127	0	0	0
TG15322	M	6	6	19,813,594	31.79	6299534	141152	4034740	8145458	0	0	0	1192710	0	0	0
TG78042-REP-L007	M	1	7	21,015,761	31.91	6705590	159170	3571993	9440885	0	0	0	1138123	0	0	0
TG200021	F	2	7	27,606,108	31.27	8633522	197096	6548658	10229714	0	0	0	1997118	0	0	0
TG15081	F	6	7	14,040,673	30.36	4262068	90706	3477579	5357849	0	0	0	852471	0	0	0
TG49452	F	6	7	23,060,658	30.90	7126214	151488	5298223	8995135	0	0	0	1489598	0	0	0
TG48481	M	2	7	25,741,520	33.41	8599981	168999	7122870	8142306	0	0	0	1707364	0	0	0
TG29291	M	1	8	17,882,945	28.78	5146543	125823	3794995	7807635	0	0	0	1007949	0	0	0
TG104612	M	2	8	25,355,530	36.10	9152575	224402	5759987	8636698	0	0	0	1581868	0	0	0
TG20481	M	6	8	14,028,781	30.61	4293535	102843	3083137	5785075	0	0	0	764191	0	0	0
TG49451	F	6	8	17,654,003	29.01	5120946	112125	5235561	5953729	0	0	0	1231642	0	0	0
TG113632	F	6	8	23,986,032	31.74	7613077	190212	6373054	8364232	0	0	0	1445457	0	0	0

Grouping variable	Av read pairs	Av % assigned
Lane 1	26,948,553	30.51
Lane 2	31,895,254	30.07
Lane 3	31,579,412	29.66
Lane 4	22,068,729	32.00
Lane 5	28,779,396	30.54
Lane 6	23,686,897	31.41
Lane 7	22,292,944	31.57

Table B4. Post-quantification QC metrics for the RNA-seq experiment

```
#!/bin/sh
#############################################################################
#script to run a fastqc batch job on a set of fastqc.gz files
#
#fastq files either need to be in input, or you need to create symbolic links
#in the input directory that point to where they are stored
#############################################################################
#input : ../input/*fastq.gz
#output : fastqc.zip containing fastqc output
#############################################################################

FILES=`find -L ../input/ -name "*.fastq.gz"`
echo $FILES

#make sure files exist before running the rest of commands
if [ "${FILES}" -eq 0 ]; then echo "no fastq.gz files found!"; exit 0; fi

mkdir ../output/FastQC

#don't unzip files, use 4 cores to process simultaneously
fastqc -o ../output/FastQC --noextract --threads=4 $FILES
#############################################################################
```

Figure B1. *Bash* scripts for RNA-seq analysis pipeline

```
#!/bin/sh
##############################################################################################
#this script will remove illumina truseq adaptors from paired end reads
#and trim low quality reads (phred <20) from the ends of reads
#
#files should be named following this convention:
#name_tag_lanenumber_R1_replication.fastq.gz
#
#for more than one sample, run qsub with -t 1-n,
#where n is number of samples sequenced, to submit as an array job
##############################################################################################
#input: ../input/*.fastq.gz
#output: ../input/*_trimmed.fastq.gz
##############################################################################################
#find the files, set up input and output names
find -L ../input/ -name "*_R1_*.fastq.gz"| sort -d -o ../input/cutadapt_file_list.txt
#make sure files exist before running the rest of commands
FILES=`cat ../input/cutadapt_file_list.txt | wc -l`
if [ "${FILES}" -eq 0 ]; then echo "no fastq.gz files found!"; exit 0; fi
echo "multiple tasks submitted"
TASKS=`cat ../input/cutadapt_file_list.txt | wc -l`
TASKS=$((${TASKS} + 1))
for (( NUM=1; NUM<${TASKS}; NUM++ ))
do
sedcommand="${NUM}p"
R1FN=`cat ../input/cutadapt_file_list.txt | sed -n "$sedcommand"`
R1BN=`basename "${R1FN}"`
R2REGEX="`echo $R1BN | cut -d'_' -f1`""*""_R2_""`echo $R1BN | cut -d'_' -f5`"
R2FN=`find -L ../input/ -name $R2REGEX`
R2BN=`basename "${R2FN}"`
R1OUTPUT="../input/""`echo $R1BN | cut -d'.' -f1`""_trimmed"".fastq.gz"
R2OUTPUT="../input/""`echo $R2BN | cut -d'.' -f1`""_trimmed"".fastq.gz"
TMP1="${SGE_TASK_ID}""_""tmp1.fastq.gz"
TMP2="${SGE_TASK_ID}""_""tmp2.fastq.gz"

echo -e "task ID: ""${SGE_TASK_ID}\n""read 1 input: ""${R1FN}\n""read 2 input:
  ↪   ""${R2FN}\n"\
"read 1 temp: ""${TMP1}\n""read 2 temp: ""${TMP2}\n"\
"read 1 output: ""${R1OUTPUT}\n""read 2 output: ""${R2OUTPUT}\n"

ADPTFW="AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC"
ADPTREV="AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT"

cutadapt -a "${ADPTFW}" --minimum-length 20 --quality-cutoff=20 \
--paired-output "${TMP2}" -o "${TMP1}" "${R1FN}" "${R2FN}"
cutadapt -a "${ADPTREV}" --minimum-length 20 --quality-cutoff=20 \
--paired-output "${R1OUTPUT}" -o "${R2OUTPUT}" "${TMP2}" "${TMP1}"

rm *tmp1.fastq.gz *tmp2.fastq.gz
done
##############################################################################################
```

```
#!/bin/sh
#############################################################################################
#script to run a fastqc batch job on a set of fastqc.gz files
#
#if fastq files are stored anywhere other than input, create symbolic links in the input directory
#############################################################################################
#input : ../input/*trimmed_fastq.gz
#output : fastqc.zip containing fastqc output
#############################################################################################

FILES=`find -L ../input/ -name "*_trimmed.fastq.gz"`
echo $FILES

#make sure files exist before running the rest of commands
if [ "${FILES}" -eq 0 ]; then echo "no trimmed_fastq.gz files found!"; exit 0; fi

mkdir ../output/trimmed_FastQC

#don't unzip files, use 4 threads in parallel
fastqc -o ../output/trimmed_FastQC --noextract --threads=4 $FILES
#############################################################################################
```

```
#!/bin/sh
#$ -cwd
#$ -j y
#$ -N bwa_align
#$ -S /bin/sh
#$ -M asaffa01@mail.bbk.ac.uk
#$ -m bea
#$ -pe mpi 8
#############################################################################
#script to align subset of reads to transcriptome using bwa
#############################################################################
export PATH="$SGE_O_PATH"

FILES=`cat ../input/bwa_file_list.txt | wc -l`
if [ "${FILES}" -eq 0 ]; then echo "no fastq files found"; exit 0; fi

mkdir ../output/bwa

if [ "${SGE_TASK_ID}" != "undefined" ]; then
echo "multiple tasks submitted"
sedcommand="${SGE_TASK_ID}p"
R1FN=`cat ../input/bwa_file_list.txt | sed -n "$sedcommand"`
R1BN=`basename "${R1FN}"`
R2REGEX="`echo "${R1BN}" | \
cut -d'_' -f1,2,3`""_R2_""`echo "${R1BN}" | cut -d'_' -f5,6`"
R2FN=`find -L ../input/ -name "${R2REGEX}"`
OUTPUTFN="../output/""bwa/""`echo "${R1BN}" | \
cut -d'_' -f1,2,3,5`""_""${SGE_TASK_ID}""_hits.sam.gz"
TEMPFN1="/scratch/""${SGE_TASK_ID}""_tmp1.fastq"
TEMPFN2="/scratch/""${SGE_TASK_ID}""_tmp2.fastq"
> "${TEMPFN1}"
> "${TEMPFN2}"
zcat "${R1FN}" | head -n4000000 > "${TEMPFN1}"
zcat "${R2FN}" | head -n4000000 > "${TEMPFN2}"
else
echo "single task submitted"
R1FN=`head -n1 ../input/bwa_file_list.txt`
R1BN=`basename "${R1FN}"`
R2REGEX="`echo "${R1BN}" | \
cut -d'_' -f1,2,3`""_R2_""`echo "${R1BN}" | cut -d'_' -f5,6`"
```

```
R2FN=`find -L ../input/ -name "${R2REGEX}"`
OUTPUTFN="../output/""bwa/""`echo "${R1BN}" | \
cut -d'_' -f1,2,3,5`""_hits.sam.gz"
TEMPFN1="/scratch/tmp1.fastq"
TEMPFN2="/scratch/tmp2.fastq"
> "${TEMPFN1}"
> "${TEMPFN2}"
zcat "${R1FN}" | head -n4000000 > "${TEMPFN1}"
zcat "${R2FN}" | head -n4000000 > "${TEMPFN2}"
fi

echo -e "task ID: ""${SGE_TASK_ID}\n""read 1 input: ""${R1FN}\n""read 2 input:
  ↪   ""${R2FN}\n"\
"output file: ""${OUTPUTFN}\n""temp 1: ""${TEMPFN1}\n""temp 2: ""${TEMPFN2}\n"

~/apps/bwa-0.7.10/bwa mem -t 8 -M -v 3 ../genome/hg19_transcriptome/mrna.fa.gz \
"${TEMPFN1}" "${TEMPFN2}" | gzip -3 > "${OUTPUTFN}"
################################################################################
```

```
#!/bin/bash
#########################################################################
#script to make tophat bowtie indexes
#########################################################################

echo "making tophat bowtie indexes:"
tophat -G ../genome/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf \
--transcriptome-index=../genome/transcriptome_index \
../genome/Homo_sapiens/UCSC/hg19/Sequence/Bowtie2Index/genome
#########################################################################
```

```
#!/bin/sh
#############################################################################################
#script to run a tophat array job on a set of fastqc.gz files
#
#fastqs should be in input folder, if data is stored elsewhere,
#need to have symbolic links in the input folder
#############################################################################################
#input : ../input/*_trimmed.fastq.gz
#output : tophat/accepted_hits.bam among others
#############################################################################################

#find the files, set up input and output names
find -L ../input/ -name "*R1_*_trimmed.fastq.gz" | sort -d -o ../input/tophat_file_list.txt
#make sure files exist before running the rest of commands
FILES=`cat ../input/tophat_file_list.txt | wc -l`
if [ "${FILES}" -eq 0 ]; then echo "no fastq files found"; exit 0; fi

echo "multiple tasks submitted"
TASKS=`cat ../input/tophat_file_list.txt | wc -l`
TASKS=$((${TASKS} + 1))
for (( NUM=1; NUM<${TASKS}; NUM++ ))
do
sedcommand="${NUM}p"
R1FN=`cat ../input/tophat_file_list.txt | sed -n "$sedcommand"`
R1BN=`basename $R1FN`
R2REGEX="`echo "${R1BN}" | cut -d'_' -f1,2,3`""_R2_""`echo "${R1BN}" | cut -d'_' -f5,6`"
R2FN=`find -L ../input/ -name $R2REGEX`
OUTPUTFN="../output/""tophat_""`echo "${R1BN}" | cut -d'_' -f1,2,3,5`""_""${NUM}"

echo -e "task ID: ""${TASKS}\n""read 1 input: ""${R1FN}\n""read 2 input:
  ↪   ""${R2FN}\n""output file: ""${OUTPUTFN}\n"

# tophat parameters
# p is number of threads to use - leave this as 8
# g is number of multihits allowed
# r is expected inner distance between mate pairs: r = library fragment size - (2 * read
  ↪   length)
# segment length = 0.5 * read length
# library type is fr-unstranded for standard illumina platforms
# G is the reference genome annotation to use
# o is output file
# need to provide location of bowtie indexes
# need to provide input fastq file
tophat -p 4 -r -10 --mate-std-dev 50 --library-type fr-firststrand -o "${OUTPUTFN}" \
--transcriptome-index=../genome/transcriptome_index \
../genome/Homo_sapiens/UCSC/hg19/Sequence/Bowtie2Index/genome "${R1FN}" "${R2FN}"
echo "${OUTPUTFN}" >> ../output/cufflinks_file_list.txt
sort ../output/cufflinks_file_list.txt -d -o ../output/cufflinks_file_list.txt
done
#####################################################################
```

```
#!/bin/sh
#$ -cwd
#$ -j y
#$ -N picard_tools_insert_size.sh
#$ -S /bin/sh
#$ -M asaffa01@mail.bbk.ac.uk
#$ -m bea
#############################################################################
#script to get insert size metrics for bwa alignment
#############################################################################
export PATH="$SGE_O_PATH"

find -L ../output/bwa/ -name "*hits.sam.gz" | sort -d -o ../output/bwa/picard_file_list.txt
FILES=`cat ../output/bwa/picard_file_list.txt | wc -l`
if [ "${FILES}" -eq 0 ]; then echo "no sam files found"; exit 0; fi

if [ "${SGE_TASK_ID}" != "undefined" ]; then
echo "multiple tasks submitted"
sedcommand="${SGE_TASK_ID}p"
R1FN=`cat ../output/bwa/picard_file_list.txt | sed -n "$sedcommand"`
else
echo "single task submitted"
R1FN=`head -n1 ../output/bwa/picard_file_list.txt`
fi

R1BN=`basename "${R1FN}"`
OUTPUTFN="../output/""bwa/""`echo "${R1BN}" | cut -d'.' -f1`"".sorted.sam.gz"
METRICS="../output/""bwa/""`echo "${R1BN}" | cut -d'.'
  ↪   -f1`""_collect_insert_size_metrics"".txt"
HISTO="../output/""bwa/""`echo "${R1BN}" | cut -d'.'
  ↪   -f1`""_collect_insert_size_metrics"".hist"

echo -e "task ID: ""${SGE_TASK_ID}\n""sam input: ""${R1FN}\n""sam output: ""${OUTPUTFN}\n"

java -Xmx2g -jar /home/saffaria/apps/picard-tools-1.124/picard.jar SortSam INPUT="${R1FN}"
  ↪   OUTPUT="${OUTPUTFN}" SORT_ORDER=coordinate
java -Xmx2g -jar /home/saffaria/apps/picard-tools-1.124/picard.jar
  ↪   CollectInsertSizeMetrics INPUT="${OUTPUTFN}" OUTPUT="${METRICS}"
  ↪   HISTOGRAM_FILE="${HISTO}"
#############################################################################
```

```
#!/bin/sh
#########################################################################
#script to run post-alignment QC
#########################################################################
find ../output/ -name "accepted_hits.bam" | sort -d -o ../output/tophat_alignments.txt
FILES=`cat ../output/tophat_alignments.txt | wc -l`
if [ "${FILES}" -eq 0 ]; then echo "no files found"; exit 0; fi

echo "multiple tasks submitted"
TASKS=`cat ../output/tophat_alignments.txt | wc -l`
TASKS=$((${TASKS} + 1))
for (( NUM=1; NUM<${TASKS}; NUM++ ))
do
sedcommand="${NUM}p"
R1FN=`cat ../output/tophat_alignments.txt | sed -n "$sedcommand"`

#set up filenames and directories
PICARDREOOUT="../"`echo "${R1FN}"| cut -d'.' -f3`"_reordered.bam"
PICARDRGOUT="../"`echo "${R1FN}"| cut -d'.' -f3`"_reordered_rg.bam"
PICARDDUPOUT="../"`echo "${R1FN}"| cut -d'.' -f3`"_reo_rg_mkdups.bam"
SORTSAMOUT="../"`echo "${R1FN}"| cut -d'.' -f3`"_reo_rg_mkdups.bam.bai"
OUTPUTDIR="`dirname "${R1FN}"`""/"
mkdir "${OUTPUTDIR}"samtools/
mkdir "${OUTPUTDIR}"picard/
mkdir "${OUTPUTDIR}"RNA-SeQC/

echo -e "task ID: ""${NUM}\n""bam input: ""${R1FN}\n" \
"reordered sam: ""${R1OUTFN}\n""outputdir: ""${OUTPUTDIR}\n" \
"picard mark dups: ""${PICARDDUPOUT}"

#add read groups to SAM for benefit of picardtools and RNA-SeQC
SAMPLE=`echo "${R1FN}" | cut -d '/' -f3 | cut -d '_' -f2`
LANE=`echo "${R1FN}" | cut -d '/' -f3 | cut -d '_' -f4`
BARCODE=`echo "${R1FN}" | cut -d '/' -f3 | cut -d '_' -f3`
RGID="P410-C5RWMACXX."${LANE}"
RGLB="TRUSEQ_dUTP_P410"
RGPL="ILLUMINA"
RGPU="C5RWMACXX-"${BARCODE}."${LANE}"

echo -e "${RGID} \n""${SAMPLE} \n""${RGLB} \n""${RGPL} \n""${RGPU} \n"
```

```
java -jar ~/apps/picard-tools-1.124/picard.jar AddOrReplaceReadGroups INPUT="${R1FN}" \
OUTPUT="${PICARDRGOUT}" RGID="${RGID}" RGSM="${SAMPLE}" RGLB="${RGLB}" RGPL="${RGPL}"
  ↪   RGPU="${RGPU}"

#reorder SAM for benefit of picardtools and RNA-SeQC
java -jar ~/apps/picard-tools-1.124/picard.jar ReorderSam INPUT="${PICARDRGOUT}" \
OUTPUT="${PICARDREOOUT}" VALIDATION_STRINGENCY=LENIENT \
REFERENCE=../genome/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa
  ↪   CREATE_INDEX=false

#mark duplicates - also required for RNA-SeQC to run
java -jar ~/apps/picard-tools-1.124/picard.jar MarkDuplicates INPUT="${PICARDREOOUT}" \
OUTPUT="${PICARDDUPOUT}" METRICS_FILE="${OUTPUTDIR}"picard/dup_metrics.txt \
REMOVE_DUPLICATES=false

#index SAM file - required for RNA-SeQC
samtools index "${PICARDDUPOUT}" "${PICARDDUPOUT}".bai

#remove intermediate BAM files
rm "$PICARDREOOUT" "$PICARDRGOUT"

#multiple metrics
java -jar ~/apps/picard-tools-1.124/picard.jar CollectMultipleMetrics
  ↪   INPUT="${PICARDDUPOUT}" \
REFERENCE_SEQUENCE=../genome/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa \
OUTPUT="${OUTPUTDIR}"picard/picard

#RNA-SeQC

#create sequence dictionary for genome fasta- required
java -jar ~/apps/picard-tools-1.124/picard.jar CreateSequenceDictionary \
REFERENCE=../genome/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa \
OUTPUT=../genome/hg19_sequence_dictionary.bam

#index genome fasta
if [ "${NUM}" -eq 1 ];then samtools faidx
  ↪   ../genome/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa;fi
java -Xmx2g -jar ~/apps/RNA-SeQC_v1.1.8.jar \
-n 1000 -o "${OUTPUTDIR}"RNA-SeQC/ -r
  ↪   ../genome/Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa \
-s ""${SGE_TASK_ID}"|"${PICARDDUPOUT}"|1" -t
  ↪   ../genome/Homo_sapiens/UCSC/hg19/Annotation/Genes/genes.gtf \
-BWArRNA ../genome/human_all_rRNA.fasta
done
#########################################################################
```

```
#! bin/sh
#$ -cwd
#$ -j y
#$ -N htseq_make_count_table.sh
#$ -S /bin/sh
#$ -M asaffa01@mail.bbk.ac.uk
#$ -m bea
#########################################################################
#script to quantify seq counts
#########################################################################
export PATH="$SGE_O_PATH"

if [ "${SGE_TASK_ID}" != "undefined" ]; then
echo "multiple tasks submitted"
sedcommand="${SGE_TASK_ID}p"
R1FN=`cat ../output/sam_name_sort.txt | sed -n "$sedcommand"`
else
echo "single task submitted"
R1FN=`head -n1 ../output/sam_name_sort.txt`
fi

R1OUTFN="..""`echo "${R1FN}"| cut -d'.' -f3`""_nsort"
OUTPUTDIR="`dirname "${R1FN}"`""/"

echo -e "task ID: ""${SGE_TASK_ID}\n""bam input: ""${R1FN}\n""reordered sam:
  ↪   ""${R1OUTFN}\n""outputdir: "
"${OUTPUTDIR}"

#sort sam by read name for HTSeq
samtools sort -n -m 1000000000 "${R1FN}" "${R1OUTFN}"

#run HTSeq to generate gene count table
samtools view -h "${R1OUTFN}".bam | \
python -m HTSeq.scripts.count --stranded=reverse - ../genome/transcriptome_index.gff > \
"${R1OUTFN}"_rev.count
#########################################################################
```

Number of DE genes called

Comparison	Model	Number of DE genes at FDR < 0.05 *limma voom*	*edgeR*	*DESeq2*
Case-control	y ~ pair + case	1	0	55
Group 2	y ~ pair + case	0	3	28
Gender	y ~ gender	156	280	287
Group 6 (mock)	y ~ mock_case	1	5	0

Comparison	Model	Number of DE genes at FDR < 0.1 *limma voom*	*edgeR*	*DESeq2*
Case-control	y ~ pair + case	1	0	75
Group 2	y ~ pair + case	0	3	48
Gender	y ~ gender	290	430	1028
Group 6 (mock)	y ~ mock_case	2	5	0

Comparison	Model	Number of DE genes at FDR < 0.2 *limma voom*	*edgeR*	*DESeq2*
Case-control	y ~ pair + case	1	3	205
Group 2	y ~ pair + case	0	6	52
Gender	y ~ gender	549	995	2176
Group 6 (mock)	y ~ mock_case	2	6	0

Table B5. Comparison of differential expression analysis methods for RNA-seq - for each regression model and for each method, the number of genes called as DE for a range of FDR thresholds is given

Proportion of expected DE genes

Comparison	Model	Proportion of expected true nulls (pi0) *limma voom*	*edgeR*	*DESeq2*	Expected number of DE genes *limma voom*	*edgeR*	*DESeq2*
Case-control	y ~ pair + case	0.83	0.88	0.82	3036	2226	3219
Group 2	y ~ pair + case	0.85	0.92	1	2520	1446	<1
Gender	y ~ gender	0.89	0.84	0.84	1889	2803	2859
Group 6 (mock)	y ~ mock_case	1	1	1	<1	<1	<1

Sensitivity and specifcity

Comparison	Model	Sensitivity *limma voom*	*edgeR*	*DESeq2*	Specificity *limma voom*	*edgeR*	*DESeq2*
Gender	y ~ gender	0.1	0.58	0.53	0.99	0.99	0.99

Comparison	Model	log(DOR) *limma voom*	*edgeR*	*DESeq2*	Fisher's exact p *limma voom*	*edgeR*	*DESeq2*
Gender	y ~ gender	1.04	2.14	2.05	2.33E-13	2.20E-16	2.20E-16

Table B6. Comparison of differential expression analysis methods for RNA-seq - sensitivity and specificity of the different methods for the male-female contrast, comparing genes called as DE with those previously known to display gender differences.

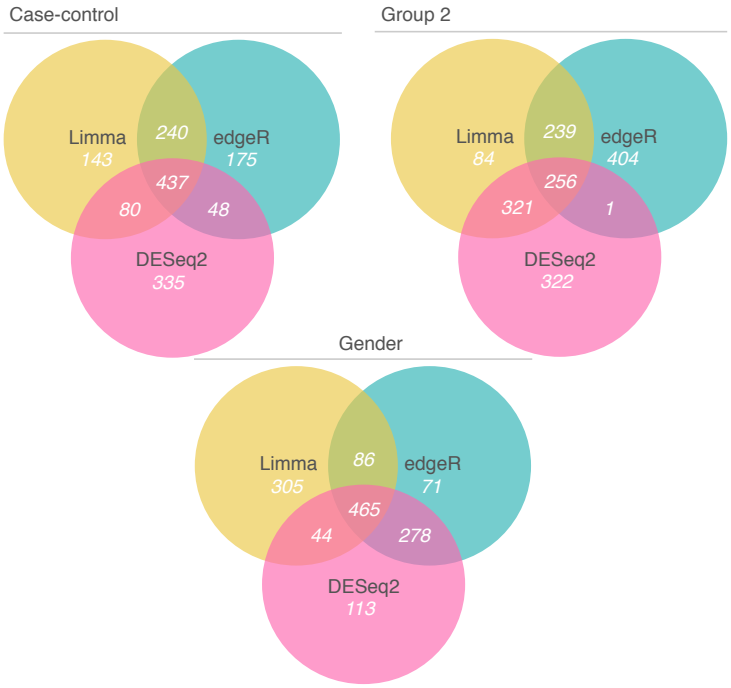

Figure B2. Comparison of differential expression analysis methods for RNA-seq - overlap between the top 900 most significantly differentially expressed genes for the three different methods for the case-control, group 2, and gender contrasts.

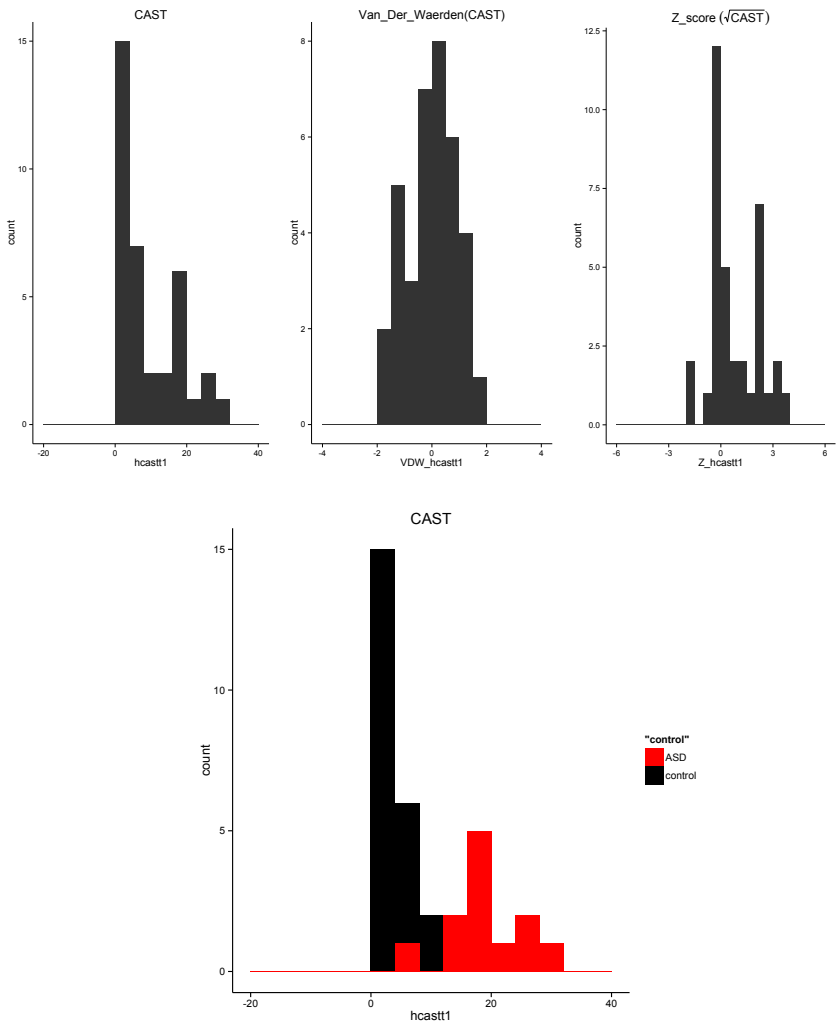

Figure B3. Distribution of CAST scores across the entire sample, along with distributions for Van Der Waerden transformed CAST scores and Z transformed CAST scores.

C Chapter 4

```
print """
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/
# Permutation analysis of methylation data:
# 1. Permutation procedure
#^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
# author. Ayden Saffari <asaffa01@mail.bbk.ac.uk>
# affiliations. Birkbeck, University of London,
#               London School of Hygiene and Tropical Medicine
#^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
# description: randomly permutes phenotypes of subjects, performs t-tests, and records max
#                            t stat value obtained for each probe over all the perms
#
# prerequisites: Python 2.7 + necessary libraries (see below)
#
# notes: colnames (subjects) in methylation data must follow this naming convention:
#                  [experimental_group]_[some other identifier]_[affected or unaffected status]
#         eg. 2_TP3_A and 2_TP3_U.
#
# input args: 1 = processor cores, 2 = permutations per thread (needs to be divisible
#                           by n cores), 3 = input methylation data (.csv)
#                            4 = experimental groups to use in analysis (0 for all subjects),
#                           5 = related samples (not implemented, set to 0),
#                           6 = equal variance (not implemented, set to 1),
#                    7 = number of permutations to generate
#                     8 = output file (.h5)
#
# output: hdf table of results, containing the p-values for each probe (rows) for each
#                   permutation (columns) and the minimum p.
#
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/
"""
import sys
import os
import subprocess
import argparse
import gc
import gzip
import time
import math
import random
```

Figure C1. *Python* code for permutations generating procedure

```
import re
import pdb
import itertools as itperm
import multiprocessing as multipro
import logging

import numpy as np
import pandas as pd
from scipy import stats as sp
import tables
import IPython

#########################
# load methylation data #
#########################

def load_data(s):
    file_name = s

    try:
        in_data = open(file_name)
    except IOError as e:
        logger.error("I/O error({0}): {1}".format(e.errno, e.strerror))

    pd.options.display.max_columns = 101
    #load dataframe
    meth_data = pd.DataFrame.from_csv(in_data, sep = ',', header = 0)
    in_data.close()
    logger.info("Dataframe dimensions: %i rows by %i columns" %(
    len(meth_data.index),len(meth_data.columns)))

    meth_data_redux = meth_data
    #sort by column label to arrange subjects by pair and phenotype
    meth_data_redux = meth_data_redux.sort_index(axis = 1)

    return meth_data_redux

########################################
# select experimental group for analysis #
########################################
```

```
#groups are numbered, all for case/control using all subjects

def sel_group(s, df):
    inp = s
    group = inp.split(",")

    if (group[0] == "all"):
        meth_data_redux_group = df
    else:
        #filter according to selected group(s)
        regex = ""
        for idx, item in enumerate(group):
            regex += "\A%s_TP*" %(item)
            if idx < len(group) - 1 :
                regex += "|"
        meth_data_redux_group = df.filter(regex = regex)

    #store size of selected group
    n = len(meth_data_redux_group.columns)
    #if group wasn't found in table or has odd number of subjects
    if (n == 0):
        try:
            raise ValueError
        except ValueError:
            print "Error: No subjects specified"
            raise

    #add row for phenotypic/case-control status
    case_control = np.empty([(n)], dtype=int)
    for idx, col in enumerate(meth_data_redux_group.columns):
        if re.match('.*_A$', col):
            case_control[idx] = 1
        elif re.match('.*_U$', col):
            case_control[idx] = 0
    meth_data_redux_group.loc["case-control",:] = case_control
    logger.info(meth_data_redux_group.loc["case-control",:])
    logger.info("Dataframe dimensions: %i rows by %i columns" %(
    len(meth_data_redux_group.index), len(meth_data_redux_group.columns)))
    return (meth_data_redux_group)
```

```
################################
# randomly permute phenotypes #
################################

#recursive function to unpack list of tuples into flat list
#(no longer required)
def unpack(a, b):
    if type(a) is not tuple:
        b.append(a)
    else:
        for i in range(0,len(a)):
            unpack(a[i], b)

#function for randomly permuting case-control labels
def rand_perm(i, df, rel):
    nperms = i

    #data paired (related samples) -> shuffle labelling within pairs
    if rel == True:
        #labels = list(df.ix["case-control", 0:2])
        labels = [1,0]
        n_labels = len(list(df.loc["case-control", :]))
        all_perms = []
        for i in range(0,int(nperms)):
            temp_perm = []
            for i in range(0, n_labels / 2):
                random.shuffle(labels)
                temp_perm = temp_perm + labels
            all_perms.append([bool(item) for item in temp_perm])

    #unpaired -> randomly assign labels by individual
    elif rel == False:
        labels = list(df.loc["case-control", :])
        all_perms = []
        for i in range(0, int(nperms)):
            random.shuffle(labels)
            all_perms.append([bool(item) for item in labels])

    #output
    unique = [list(x) for x in
```

```
            set(tuple(x) for x in all_perms)]
    logger.info("%i random permutations performed, of which %i are unique" %(
    int(nperms), len(unique)))
    return all_perms

#debug - the zeroth permutation is fixed to use the true phenotypes
def add_orig_pheno(df, pm):
    real_pheno = df.loc["case-control", :]
    temp = [bool(item) for item in real_pheno]
    pm[0] = temp
    df.loc["case-control", :] = pm[0]
    return df

####################################
# perform t-test for permuted data #
####################################

#select type of t-test to be performed
def ttest_selector(rel, var):
    if rel == True:
        t = "sp.ttest_rel(cases,controls,0)"
    else:
        t = "sp.ttest_ind(cases,controls,0,{})".format(var)
    return t

#perform for every probe in every permutation
def mp_ttest_probes_in_perm(t):
    start = time.time()
    len_df = len(mp_meth_matrix)
    pval = list()

    for perm in t:
        tup_p = tuple(perm)
        #ttest diff methylation for each probe, return vector of p-vals
        for m in xrange(0,(len_df - 1)):
            #get shuffled cases and controls from df
            #controls are those that have a 0 in the case-control row
            #select cases by taking remaining subjects
            bool_sel = np.array(perm)
            controls = mp_meth_matrix[m, bool_sel]
```

```
            cases = mp_meth_matrix[m, ~bool_sel]
            #perform ttest
            res = sp.ttest_ind(cases,controls,0)
            #add to complete list of p-vals
            pval.append(abs(res[0]))

    logger.info('%s: finished' %(multipro.current_process().name))
    #print time-taken
    elapsed = time.time() - start
    m,s = divmod(elapsed, 60)
    h,m = divmod(m, 60)
    logger.info("mp_ttest_probes_in_perm_unp: time taken: %d:%02d:%02d" %(h, m, s))
    #return complete list of p-vals (for all perms and all probes for this worker)
    logger.warn('Free memory after worker finished: {m} MB'.format(
                            m = free_memory()))
    return pval

######################################
# set up multiprocessing environment #
######################################

# set up worker pool
def mp_worker_pool(d, n):
    len_d = len(d)
    quo_task_div, rem_task_div = divmod(len_d, n)
    task_list = list()
    start_it = 0

    for i in xrange(0, n):
        if i < rem_task_div:
            inc_it = quo_task_div
        else:
            inc_it = (quo_task_div - 1)
        task_list.append(d[start_it:(start_it + inc_it + 1)])
        start_it += inc_it + 1

    return task_list

def mp_main(ts, n):
    start = time.time()
```

```
    #number of processors
    num_proc = n
    mp_pool = multipro.Pool(num_proc)

    p_results = list()
    pool_arr = mp_pool.map(mp_ttest_probes_in_perm, ts)
    p_results.append(pool_arr)

    mp_pool.close()
    mp_pool.join()

    #print time-taken
    elapsed = time.time() - start
    m,s = divmod(elapsed, 60)
    h,m = divmod(m, 60)
    logger.info("mp_main: time taken: %d:%02d:%02d \n" %(h, m, s))
    return p_results

##################
# write results #
##################

def write_file(s):
    #global pvals_tab
    global store
    #open file for writing
    file_name = s
    file_out = open(file_name, 'a')
    #write to file
    store = pd.HDFStore(file_out.name)
    file_out.close()

def mp_write_results(r, c, n):
        #get individual results back from pooled results
    start = time.time()

    #num probes
    len_r = len(r)
    #init results table, get col names and add column for min p val
    names = np.append([r[0:(len_r - 1)]],["max_t_val"])
```

```
    global pvals_tab
    pvals_tab = pd.DataFrame(np.zeros((len_r, c),dtype='float32'))
    pvals_tab.index = names

    #counter for perms
    count_p = -1
    #for each set of results from each worker
    for result in p_results:
        #for each of the lists of results
        #(multiple permutation tests all flattened together)
        for j in xrange(0, len(result)):
            pval_temp = pd.DataFrame(np.zeros((len_r, c / n),dtype='float32'))
            pval_temp.index = names
            #separate into individual perms and store as cols in results table
            for i in xrange(0,len(result[j]),(len_r - 1)):
                count_p += 1
                #store row
                temp_row = result[j][i:(i + len_r - 1)]
                #get minimum for row
                min_row = max(result[j][i:(i + len_r - 1)])
                temp_row.append(min_row)
                pval_temp.ix[:,count_p] = temp_row

            #write to csv
            start_col = j * (c/n)
            end_col = j * (c/n) + (c/n)
            pvals_tab.ix[:, start_col : end_col] = pval_temp

    #print time-taken
    elapsed = time.time() - start
    m,s = divmod(elapsed,60)
    h,m = divmod(m,60)
    logger.info("WriteResults: time taken: %d:%02d:%02d \n" %(h,m,s))

def main():
        #these variables global so that each subprocess can share same memory
    global mp_meth_matrix, all_perms, ttest, cols, p_results, pvals_tab, store
    global logger
    logger = multipro.log_to_stderr(logging.INFO)
    logger.warn('Initial free memory: {m} MB'.format(m = free_memory()))
```

```

    #parse arguments from the command line
    parser = argparse.ArgumentParser()
    parser.add_argument("numcores", type = int)
    parser.add_argument("chunksz", type = int)
    parser.add_argument("inputcsv")
    parser.add_argument("groupsel")
    parser.add_argument("relsamp", type = int)
    parser.add_argument("equalvar", type = int)
    parser.add_argument("numperms", type = int)
    parser.add_argument("outputh5")
    args = parser.parse_args()

    start = time.time()

    #load methylation data
    meth_data_redux = load_data(args.inputcsv)
    #show 1st row
    logger.info("First row of data (sorted by column/sample name) : ")
    logger.info(meth_data_redux.head(1))

    #select group for analysis
    meth_data_redux_group = sel_group(args.groupsel, meth_data_redux)
    del meth_data_redux

    #select type of t-test to use
    args.relsamp = bool(args.relsamp)
    args.equalvar = bool(args.equalvar)
    ttest = ttest_selector(args.relsamp,args.equalvar)
    logger.info("Type of t-test used: %s" %(ttest))

    #generate permutations
    all_perms = rand_perm(args.numperms,meth_data_redux_group,args.relsamp)
    meth_data_redux_group = add_orig_pheno(meth_data_redux_group, all_perms)
    mp_meth_matrix = meth_data_redux_group.as_matrix()
    row_names = meth_data_redux_group.index.tolist()
    len_df = len(meth_data_redux_group.index)
    del meth_data_redux_group
    logger.warn('Free memory after loading data: {m} MB'.format(m = free_memory()))
    num_proc = args.numcores
```

```
    num_perms = len(all_perms)

    #create hdf data store
    write_file(args.outputh5)

        #########
        # split #
        #########
    #process in chunks
    inc = args.chunksz
    logger.info("chunksize: {c}".format(c = inc))
    for n in xrange(0, num_perms, inc):
        logger.info("chunk = {m}".format(m=int(
                                (n / float(num_perms)) * int(num_perms / float(inc)))))
        #split between cores
        task_list = mp_worker_pool(all_perms[n:(n + inc)], num_proc)
        #########
        # apply #
        #########
        p_results = mp_main(task_list, num_proc)
            ############
            # combine #
            ############
            logger.warn('Free memory after all workers done: {m} MB'.format(
                                    m = free_memory()))
            mp_write_results(row_names, inc, num_proc)
        hdf_node = str(int((n / float(num_perms)) * (num_perms / float(inc))))
        logger.warn('Free memory after combining results from workers: {m} MB'.format(
                                m = free_memory()))
        store[hdf_node] = pvals_tab
        logger.warn('Free memory after writing results to h5 file: {m} MB'.format(
                                m = free_memory()))
        #cleanup
        del pvals_tab
        del p_results
        del task_list
        gc.collect()

    #close hdf data store
    store.close()
```

```

    #print time-taken
    elapsed = time.time() - start
    m,s = divmod(elapsed,60)
    h,m = divmod(m,60)
    logger.info("main: time taken: %d:%02d:%02d" %(h,m,s))

def free_memory():
        total = 0
        a = list()
        p = subprocess.Popen('free -m', shell=True, stdout=subprocess.PIPE,
                                                stderr=subprocess.STDOUT)
        for line in p.stdout.readlines():
                a.append(line)
        total = int([x for x in a[1].split()][6])
        return total

def free_memory_mac():
        total = 0
        a = list()
        p = subprocess.Popen('/usr/bin/vm_stat', shell=True, stdout=subprocess.PIPE,
                                                stderr=subprocess.STDOUT)
        for line in p.stdout.readlines():
                a.append(line)
        tempint = int(a[1].strip(' \t\n\rPagesfree:.'))
        total = ((tempint * 4096) / 1024) / 1024
        return total

if __name__ == "__main__":
        main()
```

```python
# coding: utf-8
# In[1]:

print """
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/
# Permutation analysis of methylation data:
# Subsampling procedure
#
# prerequisites: Python 2.7 + necessary libraries (see below)
# input args: 1 = processor cores , 2= input permutations file (.h5),
#                  3 = number of sample cases, 4 = output file name (.csv)
# output: .csv file
#
# author. Ayden Saffari <asaffa01@mail.bbk.ac.uk>
# affiliations. Birkbeck, University of London,
#               London School of Hygiene and Tropical Medicine
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/
"""
import sys
import os.path
import gc
import argparse
import math
import itertools as itperm
import random
import re
import time
from timeit import Timer
import multiprocessing as multipro

import numpy as np
import pandas as pd
import tables
from scipy import stats as sp


# In[2]:

########################
#load permutation data #
```

Figure C2. *Python* code for subsampling procedure

```
########################

def load_data(file_name):
    global h5
    try:
        in_data = open(file_name)
    except IOError as e:
        print "I/O error({0}): {1}".format(e.errno, e.strerror)
    h5 = pd.HDFStore(file_name)
    return h5

def select_frame(n):
    global perms
    perms = h5.select(n)
    return perms

# In[3]:

########################
# get p values         #
# for table of abs Ts  #
########################

def get_p(df,n):
    df = df.applymap(lambda x: (sp.t.sf(x,(n - 2))) * 2)
    return df


# In[165]:

def mp_sub_sample(t):
    global perms
    idx_start = t[0]
    idx_fin = t[1]

    all_res = list()

    npop = len(perms)
    nreps = 100
    inc = int((1/float(100)) * npop)
```

```
1
2       for p in range(idx_start, idx_fin):
3           print "p: %s" %(p)
4           pop = np.array(perms.ix[:,p])
5           p_res = np.zeros(101)
6           p_res[0] = 0
7
8           for reps in xrange(0,nreps):
9               np.random.shuffle(pop)
10              min_p = np.zeros(101) * np.nan
11              min_p[0] = 0
12              #for each density, get the min p of subsample starting from
13              #position 0 in the pop up to current density
14              for dens in xrange(1,101):
15                  min_p[dens] = np.amin(pop[0:int((dens * inc) + 1)])
16              #add results for this rep to the cumulative results
17              p_res += min_p
18          #divide cumulative results by reps to get mean values for each probe density
19          p_res = p_res/(float(nreps))
20          #add to list of results for each worker
21          all_res.append(p_res.tolist())
22          del min_p
23          del p_res
24          del pop
25      return all_res
26
27
28  # In[71]:
29
30  def mp_worker_pool(d, n):
31
32      quo_task_div, rem_task_div = divmod(d, n)
33      task_list = list()
34      start_it = 0
35
36      for i in xrange(0, n):
37          if i < rem_task_div:
38              inc_it = quo_task_div
39          else:
40              inc_it = (quo_task_div - 1)
```

```
        task_list.append([start_it,(start_it + inc_it + 1)])
        start_it += inc_it + 1
    print 'task list: %s' %(task_list)
    return task_list


# In[171]:

def mp_main(ts, n):
    start = time.time()
    num_proc = n
    mp_pool = multipro.Pool(num_proc)

    l_res = list()
    pool_arr = mp_pool.map(mp_sub_sample, ts)
    l_res.append(pool_arr)
    mp_pool.close()
    mp_pool.join()

    #print time-taken
    elapsed = time.time() - start
    m,s = divmod(elapsed, 60)
    h,m = divmod(m, 60)
    print "mp_main: time taken: %d:%02d:%02d \n" %(h, m, s)

    return l_res


# In[170]:

def to_table(lr, n):

    ncol = 101
    dtypes = np.repeat('f8',ncol)
    names = ["{:01d}".format(x) for x in range(0,ncol)]
    cols = zip(names,dtypes)
    tab = pd.DataFrame(np.zeros(n,dtype=cols))
    count = -1

    #get results back from each worker
```

```
    for worker in lr:
        for res in worker:
            for perm in xrange(0,(len(res))):
                count += 1
                tab.iloc[count,:] = (res[perm])
    return tab


# In[173]:

####################################
# write subsampling results to csv #
####################################

def write_file(file_name):
    global res_tab

    #open file for writing
    #file_name = raw_input("Enter name of file to save results to : ")
    #file_out = open(file_name, 'a')

    #append to file (or create new if none exists)
    if(os.path.exists(file_name)):
        file_out = open(file_name, 'a')
        pd.DataFrame.to_csv(res_tab,file_out.name,sep = ',',mode='a',header = False, index
 ↪   = False)
    else:
        file_out = open(file_name, 'w')
        pd.DataFrame.to_csv(res_tab,file_out.name,sep = ',',mode='w',header = True, index =
 ↪   False)
    file_out.close()


# In[174]:

def main():

    global h5, perms, num_proc, res_tab
    #pd.set_eng_float_format(accuracy = 8)

    parser = argparse.ArgumentParser()
    parser.add_argument("numcores", type = int)
```

```
    parser.add_argument("inputcsv")
    parser.add_argument("ncase", type = int)
    parser.add_argument("outputcsv")
    args = parser.parse_args()

    #start timer
    start = time.time()

    #hd5 file
    h5 = load_data(args.inputcsv)

    #for each dataframe in file
    for df in xrange(0,len(h5)):
        print "dataframe : %s" %(df)
        #load dataframe
        perms = select_frame(str(df))

        perms.head()
        print "Dataframe dimensions: %i rows by %i columns" %(
              len(perms.index),len(perms.columns))

        #remove last row (max t val)
        perms = perms.ix[:-1]

        print "Dataframe dimensions: %i rows by %i columns" %(
              len(perms.index),len(perms.columns))
        len_df = len(perms.columns)
        perms = get_p(perms,args.ncase)
        task_list = mp_worker_pool(len_df, args.numcores)
        l_res = mp_main(task_list, args.numcores)
        res_tab = to_table(l_res, len_df)
        write_file(args.outputcsv)


# In[175]:

if __name__ == "__main__":
    main()
```

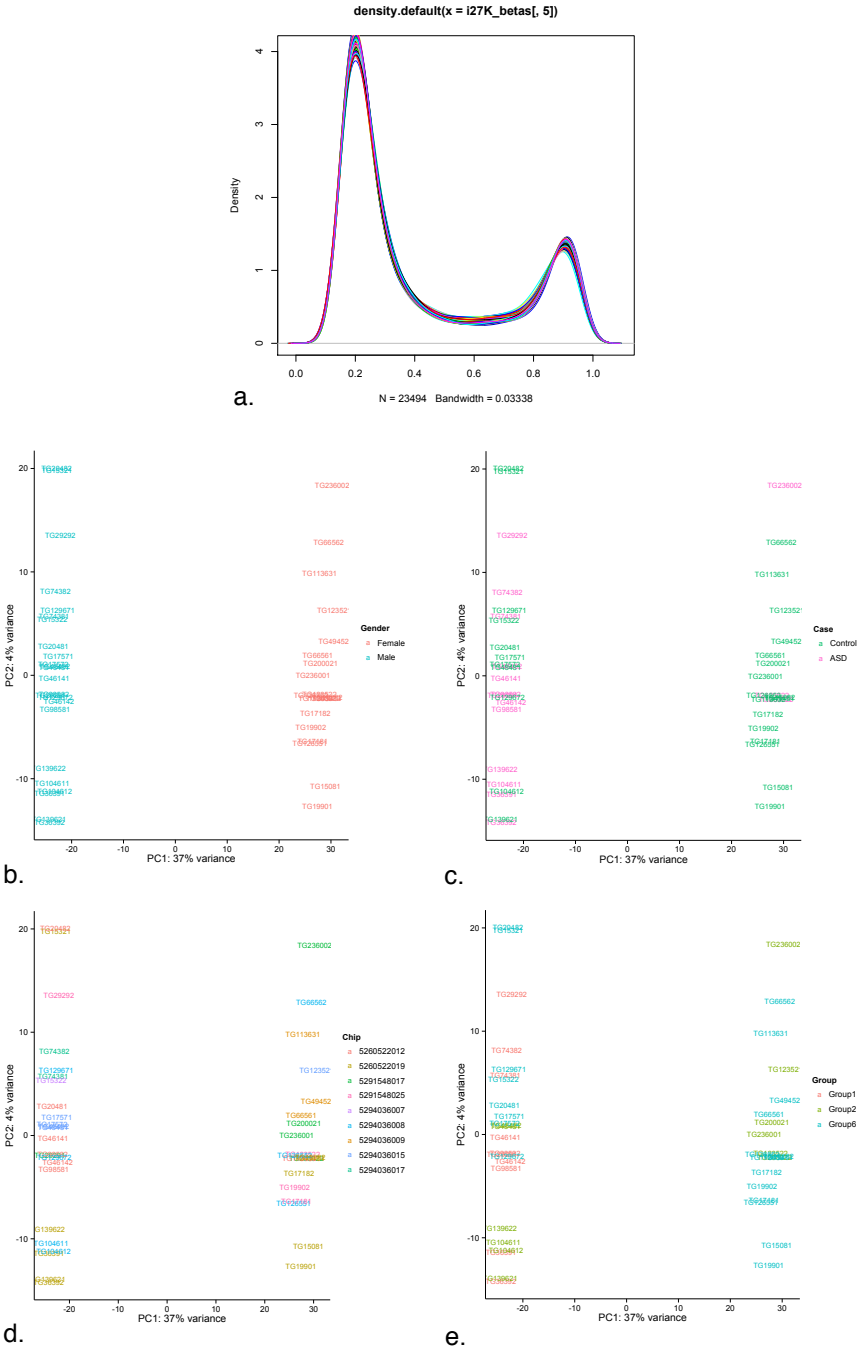

Figure D1. 27K methylation QC. a. - distribution of normalised beta values, b to e. - PCA plots of first two principal components with b. gender indicated, c. case status, d. - sentrix chip id, e. experimental group

D Chapter 5

```r
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\\/\/\/\/\/\/\/\/
# Functions for combining p-values from gene expression and methylation
# data on the same sample using Fisher's and empirical Brown's methods
#/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\\/\
#
#^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
# author. Ayden Saffari <asaffa01@mail.bbk.ac.uk>
# affiliations. Birkbeck, University of London,
#               London School of Hygiene and Tropical Medicine
#^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^#
# requires empirical Brown's method from EmpiricalBrownsMethod R package in bioconductor
#
#/\/\/\\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\/\//\/\/\/\/\/\/\/\/\/\/\/\/\/

install.packages('../R_scripts/EmpiricalBrownsMethod_0.99.1.tar.gz',
repos=NULL,type='source')
library('EmpiricalBrownsMethod')


fishers_pcomb <- function(x){
return(pchisq(-2 * sum(log(x)), 2 * length(x), lower.tail=FALSE))
}

fishers_browns <- function(meth,expr,pvals,dsets=2){

        #input:
        #meth - methylation dataset
        #expr -  expression dataset
        #pvals - matrix/dataframe containing obtained p-values for differential tests
        #for each gene for the different datasets. Should be m x d, where m is number
        #of genes tested (matching the order in data_mat), and d is the number of datasets
        #
        #dsets allows the number of datasets to be specified, default is 2 (not implemented)
        #
        #output:
        #vector of p-values combined using either empirical Brown's method, where a cross set
        #dependency exists, and Fisher's where this is not possible because gene has not been
        #measured in all the assays
```

Figure D2. *R* functions for combining p-values from gene expression and methylation data on the same sample using either Fisher's or empirical Brown's methods

```
        samples <- colnames(meth)
        n <- length(samples)
        meth_genes <- rownames(meth)
        expr_genes <- rownames(expr)
        all_genes <- rownames(pvals)
        ngenes <- length(all_genes)
        res <- vector(mode='numeric',length=ngenes)
        names(res) <- all_genes

        combi_data <- as.data.frame(setNames(replicate(n,numeric(0),
                                                       simplify = F),samples))

        for(i in all_genes){
                print(i)
                combi_data[1:2,] <- rep(1,n)

                if (i %in% meth_genes){
                        combi_data[1,] <- meth[i,]
                }

                if (i %in% expr_genes){
                        combi_data[2,] <- expr[i,]
                }

                #if a gene has no variance (because it was unmeasured in one set) brown's
  ↪   will
                #fail because covariances can't be calculated, so run Fisher's instead
                adjp <- ifelse(any(apply(combi_data,1,var) == 0),fishers_pcomb(pvals[i,]),
                                        empiricalBrownsMethod(combi_data,pvals[i,],
                                        extra_info=FALSE)[1])
                names(adjp) <- i
                res[i] <- adjp
        }
return(res)
}
```

Dataset	Sensitivity	Specificity	logDOR	Fisher's exact p
27K	0.40	0.82	0.48	4.30E-09
RNA-seq	0.31	0.86	0.44	5.43E-06

Pval combination approach	Sensitivity	Specificity	logDOR	Fisher's exact p
Fisher's	0.57	0.84	0.85	2.20E-16
Stouffer-Liptak-Kechris + Empirical Brown's	0.57	0.86	0.92	2.20E-16

Figure D3. Sensitivity and specificity for detection of gender specific genes in the gene expression and methylation datasets before and after p-value combination by Fisher's or using a combination of Stouffer-Liptak-Kechris (within methylation data) followed by Empirical Brown's (across both expression and methylation sets)