Investigating the role of microRNAs in autism

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2017

Thomas P Bleazard

Division of Population Health, Health Services Research and Primary Care
School of Health Sciences

List of Contents

Final word count: 56,794

List of Figures

List of Tables

List of Abbreviations

A-I	adenosine to inosine
ADHD	attention deficit hyperactivity disorder
ADI-R	Autism Diagnostic Interview-Revised
ADOS	Autism Diagnostic Observation Schedule
ASD	autism spectrum disorders
bp	base pairs
CLIP-seq	cross-linked immunoprecipitation
CNV	copy number variation
Ct	cycle threshold
DSM-V	Diagnostic and Statistical Manual of Mental Disorders 5
FDR	false discovery rate
GO	gene ontology
GWAS	genome-wide association study
IMGSAC	International Molecular Genetic Study of Autism Consortium
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCL	lymphoblastoid cell line
lncRNA	long non-coding RNA
mRNA	messenger RNA
MSSNG	Autism Speaks MSSNG Project
PCA	principal components analysis
PDD-NOS	pervasive developmental disorder - not otherwise specified
pre-microRNA	precursor microRNA
pri-microRNA	primary microRNA
qPCR/qRT-PCR	reverse transcription real-time polymerase chain reaction
RISC	RNA-induced silencing complex
RNA-seq	RNA sequencing
SNP/SNV	single nucleotide polymorphism/variation
UTR	untranslated region
WGS	whole genome sequencing

Abstract

Autism is a neurodevelopmental disorder involving difficulties with social communication and interaction and repetitive or restricted behaviour and interests, as well as other associated traits. Although autism is known to be highly heritable, the genetic basis of risk is still poorly understood. MicroRNAs are small non-protein coding RNAs which are highly conserved through evolution and which help regulate key developmental pathways. Previous studies showing microRNA dysregulation in autism suggest that better understanding of microRNA variants, expression and post-transcriptional regulation may help reveal underlying pathways involved in autism. MicroRNAs have also been proposed as biomarkers for better diagnosis and early intervention, leading to improved outcomes for affected individuals. In this work, we investigate the role of microRNAs in autism through small RNA sequencing of 42 autistic individuals and 10 controls, and analysis of variants in whole genome sequence data. We identified significant differential expression of 24 microRNAs, of which 5 had been previously reported as affected in autism. We developed algorithms which identified post-transcriptional regulation of microRNAs through arm usage changes, isomiRs and A-I editing. Previous work on dysregulated microRNAs hsa-miR-146a-5p, hsa-miR-132-5p and others suggest a role for Fragile X syndrome pathways in non-syndromic autism cases. We examined the set of single nucleotide variations in microRNAs in a whole genome sequencing dataset for 671 probands. We identified 101 rare SNVs within microRNA seed regions, but no *de novo* mutations. Target prediction suggested these SNVs could lead to radical changes in the set of regulated mRNAs. In addition to our study of microRNAs in autism, we also investigated research methods in this field. We found poor practice in computational analysis of microRNA functional enrichment which led to hundreds of published reports of significantly enriched pathways and functions in different diseases, conditions and biological states. We demonstrated that bias in the set of genes targeted by microRNAs in general meant that reported significant enrichments were in fact no stronger than for randomly selected microRNAs. Following concerns about poor engagement between the autism and research communities, we held workshops with 30 focus group and interview participants. We found that failure to build participatory autism research projects had led to reduced trust, recruitment problems, skewed research goals and poor dissemination. We developed recommendations for improved practice in collaboration with the autism community. Together, our work provides a framework for improved practice in both autism research and microRNA bioinformatic analysis. Our analysis of microRNAs in autism through small RNA sequencing data reveals important dysregulated pathways and contributes to a better understanding of microRNA post-transcriptional regulation in autism.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in the University's policy on Presentation of Theses.

Acknowledgements

I would like to thank my supervisors Sam Griffiths-Jones and Janine Lamb for all their efforts and help with this thesis. I have been lucky to have them to support me for four years. I thank my labmates for their friendship and support during my project, for answering all manner of questions and being always open-minded and thoughtful. I would also like to thank my family, my girlfriend, my housemates and friends for their kindness and understanding during my PhD.

I thank the autism community participants of our Building Bridges workshops, individuals in the IMGSAC cohort, MSSNG cohort and control cohorts for taking part in this research. I thank the IMGSAC team and MSSNG project members for providing samples and data for this work. I thank the Medical Research Council and Autistica for their funding to support this research.

Thesis Outline

This thesis contains an introduction (chapter 1), four manuscripts (chapters 2 to 5), and a final concluding section (chapter 6). This work is presented in journal format due to a branching out from the core theme of microRNA in autism to participatory autism research, rare variants in whole genome sequencing data, and critical assessment of computational methods for microRNA analysis.

In chapter 1, we provide a general introduction to research on microRNAs in autism. Chapter 2 presents the central work of this thesis, analysis of small RNA sequencing data for the identification of microRNA variants, differential expression and post-transcriptional regulation. We also report validation experiments for variants and differential expression. Chapter 3 is a manuscript in preparation, examining microRNA variants in a whole genome sequencing dataset. Chapter 4 is a minimally revised version of published material (Bleazard, Lamb, & Griffiths-Jones, 2015). It presents an analysis of microRNA functional enrichment analysis and demonstrates bias in a common method. Chapter 5 is a manuscript in preparation in which we report on workshops, focus groups and interviews with the autism community. This manuscript is written in consultation with autistic participants. Finally, chapter 6 provides a general discussion and overview of the key findings presented in this thesis.

Contributions

All studies in this thesis except chapter 5 are authored by Thomas Bleazard (TB), Sam Griffiths-Jones (SGJ) and Janine Lamb (JL). The studies were jointly conceived and planned, and manuscripts were jointly written. TB performed all experiments, analysed all data and drafted the manuscripts.

RNA extraction and quality control for autism and control samples for chapter 2 was performed by Eleni Tsitsiou. PCR for variant validation in chapter 2 was performed by TB and Victoria Coyne.

Chapter 3 makes use of data generated by the MSSNG Project, including the Hospital for Sick Children (Toronto) and Autism Speaks Inc.

Chapter 5 was authored by TB, Anat Greenstein (AG), Daniel Poole (DP), Peter Baimbridge (PB), Karen Stevenson (KS), James Pelham (JP) and Emma Gowen (EG). The studies were jointly conceived and planned by TB, AG, DP, PB and EG. Workshop focus groups were held by TB, AG, DP and EG. Thematic network analysis was performed by TB and AG. Autistic participants KS and JP were consulted on manuscript themes. The manuscript abstract, introduction, focus group results except power relations sections, and discussion were drafted by TB. The manuscript methods and questionnaire results were drafted by DP. Manuscript power relations sections were drafted by AG. The final manuscript was jointly written by all authors.

1. Introduction

In this introduction, we will cover the background to a study of microRNAs in autism. We first introduce microRNAs in general, giving basic findings on their biogenesis and functional role in regulating gene expression. We then move to a broad overview of autism, with the current understanding of its diagnosis and classification, phenotypic features, and genetic basis. We address controversies with autism research and provide a brief justification for approaches and terminology used in this project. We provide an introduction to known roles of microRNAs in diseases, and especially recent discoveries on involvement in neurodevelopmental disorders and autism. We survey some of the computational methods used to link microRNAs to disease and outline some critical issues with these methods. We introduce the investigative methods used in our project, describing use of RNA sequencing in particular. Finally, we present the aims of this PhD project, and an overview of the research undertaken.

1.1. Autism

Autism is a neurodevelopmental disorder whose symptoms manifest before the age of three and remain throughout life. It is characterised by impaired communication, difficulties with social interaction and restricted or repetitive behaviour. Classic autism was previously considered a member of the wider group of autism spectrum disorders (ASD), including Asperger syndrome where language and cognitive development are not delayed, and pervasive developmental disorder – not otherwise specified (PDD-NOS) (Lai, Lombardo, & Baron-Cohen, 2013). In the recent Diagnostic and Statistical Manual of Mental Disorders fifth edition (American Psychiatric Association, 2013), these three sub-types are subsumed by the general term autism spectrum disorder. Boundaries for ASD are also drawn more sharply (Baker, 2013), to require "persistent deficits in social communication and social interaction..." and "restricted, repetitive patterns of behaviour, interests or activities."

1.1.1. Cognitive theories of autism

There are three main cognitive theories which attempt to explain autism (Rajendran & Mitchell, 2007). The 'theory of mind' explanation is based on the failure of many autistic people to infer others' mental states effectively. 'Theory of mind' in a narrow sense means the idea that agents possess subjective beliefs and intentions. More generally, the theory of mind may include decoupling the real from the imaginary, failure of which could explain lack of pretend play in autistic children. Development of a theory of mind by a child requires complex cognitive mechanisms, damage of which would likely lead to social and language deficits as in autism. However, such a model does not explain why autism symptoms are observed in children before they reach the age where theory of mind is developed.

Autism very often presents together with sensorimotor disorders, inspiring the central coherence theory of autism (Levy, 2007). This explains, for example, why autistic children are better at spotting hidden figures by a preference for attention to detail rather than drawing together to build a broad picture. This would also better account for repetitive and stereotyped behaviour in autistic people. The Bayesian brain model also provides a promising perspective on sensorimotor peculiarities (Seriès & Seitz, 2013). In this model of perception as Bayesian inference, attenuation of prior expectations could lead to the excessive attention to detail observed in autism, since without well-defined prior information, inference commonly overfits data (Pellicano & Burr, 2012). This theory also links well with the theory of mind hypothesis, in that the central problem is a failure to build mental abstractions.

The theory of executive dysfunction originated with the observation that autism often mirrored specific brain injury patients. Executive function includes planning action, decision-making, judgment and self perception (Rajendran & Mitchell, 2007), which tend to be affected in autistic people.

A problem with the executive dysfunction theory is that similar deficits are reported in different diseases, such as schizophrenia and obsessive compulsive disorder.

As yet, none of the psychological theories of autism have been entirely convincing, which motivates the search for dysfunction at a deeper biological level. An interesting aspect of autism is that it involves abnormalities to some important distinguishing features of humans – social cognition and language – and so, despite the difficulty, understanding autism may help us to understand humans more generally (Geschwind, 2011).

1.1.2. Prevalence and diagnosis

Estimating the prevalence of autism in the population is complicated by sensitivity to different methods of diagnosis. A recent Korean study of the 7-12 year-old population of a village classified 2.64% of children as positive for ASD (Y. S. Kim et al., 2011). In general, rates of diagnosis have increased over time. However, these changes may be due to changing definitions, tests and awareness (Baker, 2013). For example, the term 'autism' has changed radically in its meaning over the last century (Evans, 2013), moving from an association with excess imagination to its deficit. Modern diagnosis may be done by the Autism Diagnostic Interview-Revised (ADI-R) followed by the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000). The former has the format of an interview with the parents of a child, where interviewers score responses to questions about the child's behaviour. The ADOS is commonly used with the ADI-R, and takes the form of a protocol of tasks involving interaction between the examiner and child. The 3di is an alternative published diagnostic tool which specifically measures intensity of different autism features (Skuse et al., 2004). The structure of the ADI-R has been explored previously, showing an apparent structure with two factors for severity of social interaction effects and repetitive and restricted behaviours (Snow, Lecavalier, & Houts, 2009).

1.1.3. Subclassification

There have been several attempts to further subdivide the heterogeneous autistic spectrum by detailed phenotype or gene expression, to reduce the substantial variability within groups (V. W. Hu, 2013). One study selected certain ADI-R item scores and used them as quantitative traits for genome-wide association studies (GWAS), for example (V. W. Hu & Steinberg, 2009). The same group also ran clustering algorithms on the combined feature space of ADI-R item scores, which allowed demarcation of four subgroups with phenotypic clusters (V. W. Hu & Steinberg, 2009). Such an approach has promise in helping to reduce noise from heterogeneity in underlying biology, and perhaps potential for development of personalised treatment of autism analogous to that of cancer. One question raised by the great genetic heterogeneity is how it can lead to convergent phenotypes. A helpful approach to this problem is to track the effect of genetic factors in derailing normal developmental processes (Jones & Klin, 2013).

1.1.4. Comorbidity

A notable feature of the autistic phenotype is comorbidity with other diseases and conditions. These include intellectual disability (affecting ~45% of autistic individuals), attention deficit hyperactivity disorder (ADHD), epilepsy (affecting 8-30%), gastrointestinal problems, immune dysregulation, genetic syndromes in syndromic autism (5% of autism cases), sleep disorders (affecting 50-80%), anxiety (affecting 42-56%), depression, obsessive-compulsive disorder (7-24%) and psychotic disorders (12-17%) (Lai et al., 2013). Aggressive behaviours, schizoid personality disorder, self-injurious behaviours and pica are also common. Some comorbidities in autism suggest shared developmental pathways. For example, comorbidity with intellectual disability and epilepsy may reflect general dysregulation in neural development (Gilissen et al., 2014). Other comorbidities such as anxiety and depression may be understood as responses to the experience of autism (Milton, 2012). A better understanding of the interaction of autism

with syndromes such as Fragile X syndrome and comorbid conditions may lead to broader avenues for treatment in autism (Gotham et al., 2015). Studies of genes involved in schizophrenia and intellectual disability identify significant overlap with autism (Kenny et al., 2014).

1.2. The genetics of autism

1.2.1. Genetic heritability

A key topic in the epidemiology of autism is the relative contribution of environmental and genetic factors to population variation. A standard approach to estimation of trait heritability is to measure different rates of phenotypic concordance for monozygotic and dizygotic twins. Studies using this approach in ASD have reached different estimates for genetic heritability. A recent study found a heritability of 38% (Hallmayer et al., 2011), while other estimates have been as high as 75% (Colvert et al., 2015) and 90% (Bailey et al., 1995). While these estimates imply that environmental factors deserve attention, there are few strong candidates for investigation (Landrigan, 2010). Known environmental factors causing autism include complications during pregnancy such as viral infections, exposure to valproic acid, and apparently minor effects from air pollution and pesticides (Lai et al., 2013). Other correlations with autism risk such as paternal age and higher rates in cities with more information-technology jobs may reflect underlying genetic risk, through *de novo* mutation and assortative mating and differences in technical talent respectively (Lai et al., 2013). Estimates suggest autism affects approximately 2-3 times more males than females, although it is possible that females with autism are under-diagnosed. Studies comparing single nucleotide polymorphism and single nucleotide variation (SNP and SNV) burdens and concurrent behavioural or cognitive problems show females with autism diagnosis are affected by significantly more than males (Gilman et al., 2011; L. Liu et al., 2013; Zhao et al., 2007).

1.2.2. Common variants

Attempts have been made to find common genomic variants contributing to susceptibility to autism by large genome-wide association scans (Anney et al., 2012). In these studies, a large number of cases and controls are genotyped by single nucleotide polymorphism (SNP) array, and then the genome is searched for SNPs associated with disease. The GWAS approach struggles to find variants with very low effect sizes, since large sample sizes are required for sufficient power (IMGSAC, 2001). In particular, studies using 4,305, 4,233 and 1,369 probands for discovery have previously found significant loci respectively at 5p14.1, 5p15.2 and 20p12.1 (Devlin, Melhem, & Roeder, 2010). However, combined analysis of these data would not support any significant loci, and indeed a larger meta-analysis of multiple autism genotype datasets including 16,539 probands did not find any significant loci (ASDWGPGC, 2017). Use of statistical methods to estimate undiscovered loci based on previous studies suggest that studies using tens of thousands of probands are required to discover SNPs with weak effects, for example finding SNPs with odds ratios of 1.15 with 0.95 power was estimated to require 16,489 cases (Devlin et al., 2010). The approach of GWAS may also fail to find variants which are poorly tagged by the linkage disequilibrium structure in the population, for example very rare or *de novo* mutations (Yuen et al., 2016). Despite the problem of very low separate effect sizes, it is possible to estimate a combined effect of all tagged common variants together by covariance between phenotype and (genome-wide) genotype (Yang et al., 2010). Application of this method to autism gave a lower bound for additive genetic heritability of 60% for multiplex (more than one proband) and 40% for simplex (single autistic child) families (Klei et al., 2012). Interestingly, genes implicated in autism have overlap with those for schizophrenia, ADHD, epilepsy and intellectual disability, reflecting common comorbidity of these conditions. However, an alternative approach investigating common SNP-based coheritability between ASD and other diseases showed modest similarity with schizophrenia and negative coheritability with ADHD (S. H. Lee et al.,

2013).

1.2.3. Rare variant models of susceptibility

Much attention has been focused on rare and *de novo* variants in autism in recent years. This is partly motivated by the difficulties of finding common variants through linkage scans and GWAS, and also by models of risk that account for simplex and multiplex family differences through *de novo* mutation inheritance (Zhao et al., 2007). In the *de novo* mutation hypothesis, the different concordance rates for mono- and dizygotic twins and siblings are explained by parental germline mutations affecting various risk genes. This model predicts that *de novo* single nucleotide variations (SNVs) will be observed in unaffected mothers (for unknown reasons, penetrance of the autism phenotype is much lower in females (Gilman et al., 2011)) whose male children then have high risk for autism. Indeed, in high-risk families (where the first two children born have autism) the third child has ~50% risk, which is very similar to highly penetrant dominant transmission. Fitting parameters to the data, the simple risk model described above fitted well with observed family rates (Zhao et al., 2007). Alternative hypotheses for the genetic architecture include major effect rare variants plus a sensible contribution from a background of common variants (supported by the large combined contribution inferred for common variants) and 'two hit' rare variant models (Berg & Geschwind, 2012).

1.2.4. Discovery of rare variants in autism

A review of the literature investigating rare variants in autism suggests that approximately 20% of ASD cases can be explained by copy number variations (CNVs), syndromic autism including fragile X syndrome, and single gene and metabolic disorders (Berg & Geschwind, 2012). Rare genomic variants causing disease may be effectively discovered by whole exome sequencing, if they lie within captured regions (Buxbaum et al., 2012). Those genes commonly deleteriously affected by such variants in

cases, or otherwise providing strong evidence for a causal role, may then be considered autism risk genes. Similarly, following a hypothesis that a subset of autism is caused by rare disruption of both homologous copies of key genes, whole exome sequencing was used to look for homozygous loss-of-function variants in a recent study (Lim et al., 2013). Twice as many complete gene knockouts were found in autistic people, and the combined effect of this was estimated to cause 3% of cases. Studies following a model of rare variants in autism are able to identify more specific biological processes and pathways implicated in particular cases, for example loss-of-function variants in synaptic genes (Kenny et al., 2014), in comparison with GWAS producing broader findings, such as overlap with schizophrenia common variation (ASDWGPGC, 2017). The genes discovered by such approaches, as well as by linkage scans, are catalogued by the Autism Database (AutDB), which lists genes with varying levels of confidence (Basu, Kollu, & Banerjee-Basu, 2009).

1.2.5. Mouse models of autism

Knockout mouse experiments can provide strong validation of actual causality of genes in autism. Although there are obvious reservations about the similarity of mouse and human brains and applicability of social behaviours, studies have been able to replicate human phenotypes in the mouse by knocking out genes identified previously. This includes vocalization (mirroring language effects in humans), restrictive and repetitive behaviour and social effects in *Cntnap2* (Peñagarikano et al., 2011), *Nlgn4*, *En2*, *Gabrb3*, *Oxt*, *Avpr1b* and *Fgf17* knockouts (Berg & Geschwind, 2012).

1.2.6. Affected pathways

Rare variants are best studied together by analysis of the pathways they affect (Gilman et al., 2011). Such studies suggest that autism genes are involved in regulating or are regulated by neuronal activity; some genes are

to do with activity-dependent protein metabolism at the post-synaptic density (the area where receptors are gathered at a synapse); and some genes work in neuronal cell adhesion (binding a cell to its neighbours) (Berg & Geschwind, 2012) as shown in Figure 1.1.

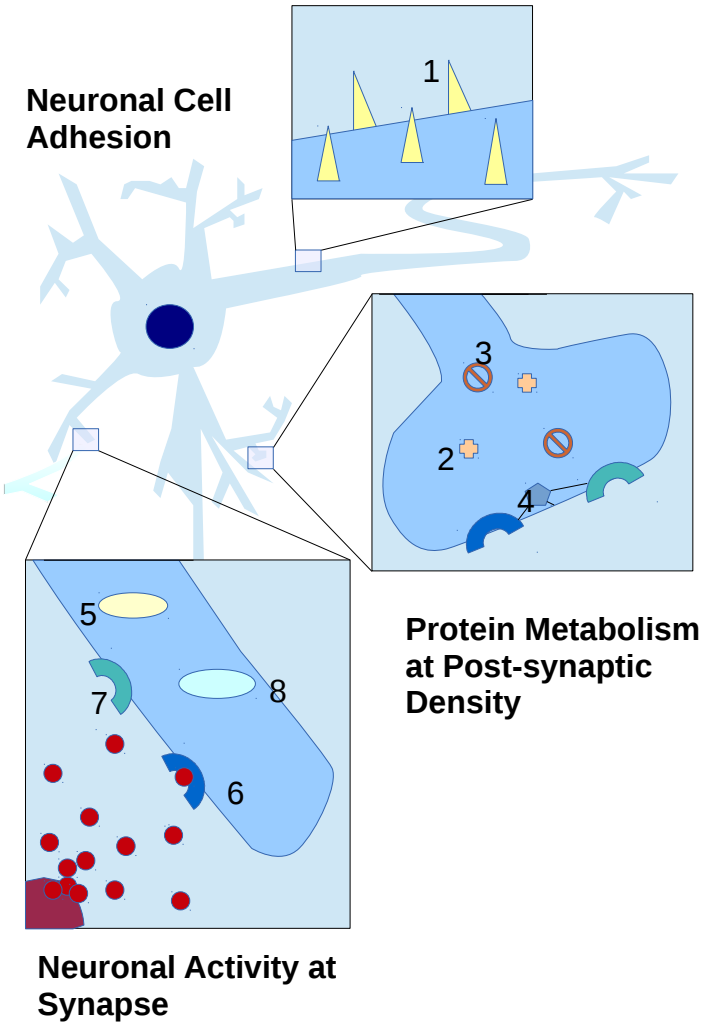


Figure 1.1. Selected autism genes functioning in the neuron

(1) *CNTN4*, *CNTN6*, *NLGN1-4*, *NRXN1*, *PCDH9*, *CHL1* and *CNTNAP2* involved in adhesion and neurite outgrowth. (2) FMRP and associated proteins regulate protein synthesis. (3) *UBE3A*, *PARK2*, *RFWD2*, *FBXO40* and *USP7* involved in protein degradation. (4) Scaffolding proteins coded for by *SHANK2* and *SHANK3*. (5) Genes *SCN2A* and *SCN1A* code for sodium channel subunits. (6) *GRIN2B* and *GRIN2A* code for NMDA receptor subunits. (7) *GRIK2* and *GRIA3* code for other glutamate receptor subunits. (8) *CACNA1C* and *CACNA1H* code for calcium channel subunits.

Genes implicated previously that are known to regulate neuronal activity include *SCN2A*, *SCN1A*, *GRIN2B*, *GRIN2A*, *GRIK2*, *GRIA3*, *CACNA1C* and *CACNA1H*, which code for ion channels and NMDA receptor subunits. Several genes whose transcription depends on neuronal activity were also implicated, including *UBE3A*, *DIA1*, *PCDH10*, *NHE9*, *UBE3B*, *CLTCL1*, *NCKAP5L* and *ZNF18* (Berg & Geschwind, 2012). Changes to these genes may impact synaptic plasticity, for example, which could lead to autistic phenotypes. Genes involved in protein metabolism at the post-synaptic density include *FMR1*, in which repeat expansion results in absence of FMRP, which is supposed to regulate protein synthesis at the synapse, leading to fragile X syndrome. Other implicated genes which are involved with FMRP include *CYFIP1*, *MET*, *PTEN*, *TSC1*, *TSC2* and *NF1*, which are part of a pathway involved in cellular proliferation. CNV studies also show affected genes involved in protein degradation at post-synaptic densities, including *UBE3A*, *PARK2*, *RFWD2*, *FBXO40* and *USP7*. Also affected are scaffolding proteins at post-synaptic densities, *SHANK2* and *SHANK3*. These genes again permit a straightforward explanation for pathophysiology, based on problems in neuronal signalling and plasticity (Berg & Geschwind, 2012). Several genes implicated in autism are involved in neuronal cell adhesion, including *CNTN4*, *CNTN6*, *NLGN1-4*, *NRXN1*, *PCDH9*, *CHL1* and *CNTNAP2*. The last of these is particularly interesting due to its appearance in a syndromic form of autism and ability to generate autistic phenotypes when removed in mice (Peñagarikano et al., 2011). Again, this points towards problems with neural connections and general structure in the brain.

1.2.7. Autism transcriptomics

Beyond genomic variants, attempts have been made to investigate the autism transcriptome. In one study, RNA sequencing (RNA-seq) was applied to post-mortem brain tissue in 19 cases and 17 controls (Voineagu et al., 2011). To understand the networks of genes that were perturbed, weighted-gene co-expression network analysis was used. This split the

genes into modules, with a representative eigengene for each that predicted the activity of genes within the module. Differential expression of the eigengenes between cases and controls was then calculated. The authors found a neuronal module enriched for known susceptibility genes and an immune and inflammatory response module both differentially expressed in autism. Interestingly, the neuronal module was enriched for autism GWAS signals, while the immune response module was not, suggesting that only the neuronal module was genetically mediated. Other studies of the transcriptome in autism have shown differential expression of immune-related genes in the brain (Garbett et al., 2008; Gupta et al., 2014), neuronal activity-dependent genes (Gupta et al., 2014), and long non-coding RNA (lncRNA) dysregulation. By mapping known autism-related genes to coexpression networks based on transcriptome sequencing data, it has been suggested that FMRP may be a major connector for affected genes (Parikshak et al., 2013). Following a similar approach with multiple brain expression datasets from foetus to adult, it was suggested that midfetal cortical projection neurons may be a point of convergence for networks containing known disrupted genes (Willsey et al., 2013).

1.3. Community concerns about autism research

Autism is a particularly controversial subject, and there are some serious issues regarding autism research and interaction with the community. Most notably, there are controversies surrounding autism cures and eugenics, and autism is still particularly poorly understood and treated in society (Bagatell, 2010). The history of 'refrigerator mothers' theory and psychoanalytic approaches remains relevant to experiences of autistic people today (Chamak, 2008). These problems have led to long-term and damaging disconnects with the autism community (Pellicano, Dinsmore, & Charman, 2014). Several of these issues and controversies are pertinent to the research presented in this thesis, with use of genetic data in the MSSNG project (Yuen et al., 2015), diagnostic approaches, and potential of microRNAs for use as biomarkers (Walsh, Elsabbagh, Bolton, & Singh, 2011).

1.3.1. Terminology in this work

There have historically been fundamental problems with describing autism. For example, technical descriptions in the DSM-V provide circular references to 'clinical significance' (Spitzer & Wakefield, 1999). Research papers produced by various groups have referred to slightly different conditions across geographical regions and time periods (Lai et al., 2013), meaning it is unclear, for example, to what extent changes in autism diagnosis rates are due to changes in diagnostic practices versus actual changes in prevalence (Y. S. Kim et al., 2011). Terminology is also important as it can be taken to imply inferiority or a one-sided view of autism (Milton, 2012).

In an attempt to encompass references to studies with different samples, we refer to 'autism' for samples in our International Molecular Genetic Study of Autism Consortium (IMGSAC) cohort based on the application of ADI-R and ADOS for diagnosis (IMGSAC, 2001), and we use 'autism spectrum disorder' where studies use diagnosis based on ASD as defined by the DSM (American Psychiatric Association, 2013). We use the term 'neurodevelopmental disorder' to reflect changes to normal functioning, without negative connotations about the person affected. We present microRNAs as potential 'biomarkers' for autism, as future postnatal diagnostic methods. The literature on community engagement also argues that better practical outcomes can be achieved from successful autism research if dissemination is improved (Elsabbagh, Yusuf, Prasanna, Ruff, & Fehlings, 2014). Therefore, we will explore using a more collaborative approach to writing on our work in community engagement, where decisions on what areas to highlight and conclusions are based on compromise and dialogue (Chown et al., 2017).

1.3.2. Participatory research

A 2014 review of participatory partnerships found only seven published studies in autism and other neurodevelopmental disorders (Jivraj, Sacrey, Newton, Nicholas, & Zwaigenbaum, 2014).We will therefore present as part of this thesis work in community engagement and recommendations for improved collaboration with the autism community. This effort is based on previous reports on participatory research in learning disabilities, and recommendations for flexible and inclusive approaches (McClimens, 2008; Nind & Vinha, 2012). This work showed that inclusive research is still undeveloped, and that exploration of effective methods for better working together would be valuable for researchers.

1.4. MicroRNA

Few studies have considered the possible role of non-protein coding genes and transcription in autism (Hicks & Middleton, 2016). MicroRNAs are a species of small RNA found in eukaryotes that work to fine-tune mRNA abundance and translation. There are 1,872 mature microRNAs annotated by miRBase v20 in humans (Kozomara & Griffiths-Jones, 2011). MicroRNAs are highly conserved and are thought to target over half of the genes in the genome (Bartel, 2009). As such, they are important regulators of expression across almost all pathways, able to effect widespread control of the transcriptome. Further, because microRNAs affect such a large number of genes when perturbed, they are of great interest in the study of disease and neurodevelopmental disorders.

1.4.1. MicroRNA biogenesis

There has been much research on the biogenesis of microRNAs, with a clear picture emerging for canonical production (Figure 1.2).

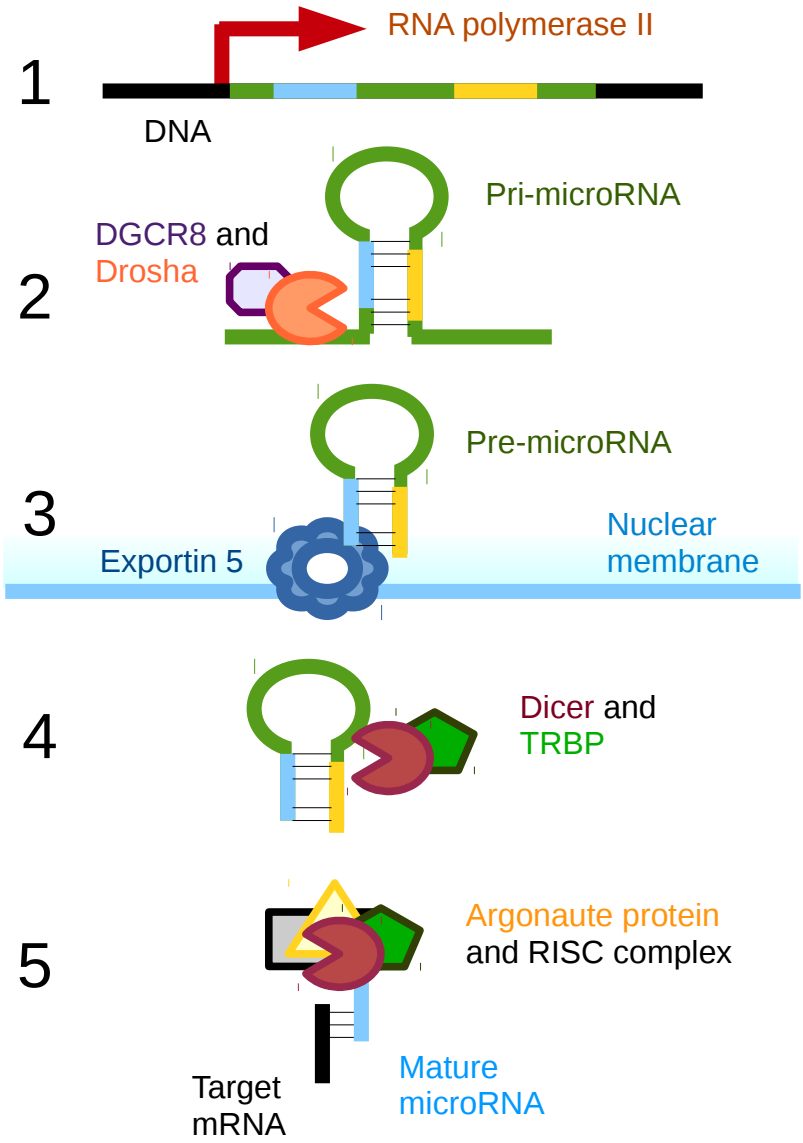


Figure 1.2. The canonical pathway for microRNA biogenesis

(1) Pri-microRNA is transcribed from the genome; (2) the base of the hairpin is cleaved by Drosha; (3) the pre-microRNA is exported to the cytoplasm by Exportin 5; (4) the loop of the hairpin is cleaved by Dicer; (5) the mature microRNA guides the RNA-induced silencing complex (RISC) in silencing a target mRNA.

MicroRNAs are initially transcribed from the genome by RNA polymerase II (pol II) (Krol, Loedige, & Filipowicz, 2010). About 30% of microRNAs in animals are clustered into genomic blocks from which they may be transcribed together (Marco, Ninova, Ronshaugen, & Griffiths-Jones, 2013). Transcription results in the primary microRNA, or pri-microRNA, which is a long strand of RNA that includes hairpins formed by complementary base-pairing. Still inside the nucleus, a protein complex of Drosha and DGCR8 binds and cuts the hairpin (or hairpins in the case of co-transcribed clusters). This produces the pre-microRNA hairpin ready for export to the cytoplasm. Some pri-microRNA come from the introns of protein-coding genes and lncRNAs, and a small proportion of introns contain 'mirtrons', which when spliced produce pre-microRNA, skipping processing by Drosha entirely. The pre-microRNA are then exported out of the nucleus by the protein Exportin 5. In the cytoplasm, the protein Dicer, assisted by TRBP (the human immunodeficiency virus transactivating response RNA-binding protein), removes the loop of the pre-microRNA to leave a ~20 bp mature microRNA duplex (Krol et al., 2010). The argonaute protein AGO2 may also assist Dicer by cleaving the 3' end of some pre-microRNAs. In the case of at least one microRNA (mir-451), processing is entirely independent of Dicer, with AGO2 cleavage and 3' exonucleic trimming producing the mature microRNA (Cheloufi, Dos Santos, Chong, & Hannon, 2010). Finally, from the mature duplex, one of the strands is discarded and one strand is incorporated into the RNA-induced silencing complex (RISC), of which the protein AGO2 is also a key component.

1.4.2. Silencing of mRNA

The microRNA in RISC acts as a guide, usually by imperfect base-pairing near its seed region (nucleotides 2 to 8) with a matching 3' UTR (untranslated region) sequence in a target mRNA. Each microRNA therefore has a set of genes which it is able to regulate based on its ability to bind successfully. There exist various models for how and when this binding will occur. Basic models require a simple matching of the seed region, to which

consideration of the free energy of hybridisation for proximal bases may be added (Enright et al., 2003; Krüger & Rehmsmeier, 2006). In some known microRNA-mRNA target pairs, the interaction has been confirmed as dependent on such seed pairing using constructs with mismatching bases (Doench & Sharp, 2004). Other models also take into account the energy required for target site accessibility (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007) as well as allowing for binding in the protein coding sequence (CDS) of some mRNAs (Reczko, Maragkakis, Alexiou, Grosse, & Hatzigeorgiou, 2012). There are also experimental approaches to identify microRNA binding sites genome-wide, through crosslinking and immunoprecipitation (CLIP-seq) (J.-H. Li, Liu, Zhou, Qu, & Yang, 2014). Luciferase reporter assays are also used to confirm microRNA-mRNA interaction (Q. Yan et al., 2013). In general, evolutionary conservation and conserved seed pairing are strong indications of true binding. However, computational tools using these methods tend to produce very divergent target predictions (Ritchie, Flamant, & Rasko, 2009).

After successful binding of a microRNA to an mRNA target binding site, silencing is then achieved through RNA degradation and translational repression. In plants, and less commonly in animals, an Argonaute protein cleaves the mRNA (Pasquinelli, 2012). In animals, the microRNA-induced silencing complex (RISC) can act to induce degradation of the bound mRNA promoted by deadenylation in which AGO, GW182 and PABP proteins participate. By removing the mRNA poly(A) tail, this process leads to exonucleolytic degradation of the mRNA (Fabian et al., 2011). Silencing can also be achieved through repression of translation. This process is less fully understood, but there is evidence for proteolysis of the nascent peptide, and inhibition of translational initiation and elongation (Pasquinelli, 2012). The 3' UTRs of target genes are evolutionarily conserved to maintain this system of regulation (Friedman & Farh, 2009). The targeting of genes by microRNAs is frequently used in regulatory networks, for example in a timer system for differentiation of neural progenitor cells (Bonev, Stanley, & Papalopulu, 2012).

1.4.3. MicroRNA regulation

There are various aspects of microRNA regulation that may impact their activity, including regulation of transcription, regulation of processing and generation of isomiRs (different mature sequences generated from the same microRNA gene), arm switching, microRNA editing and imprinting. At the root of microRNA transcription, regulation by transcription factors is similar to that of protein-coding genes. Changes to the machinery processing primary transcripts have been observed in cancer, along with global reduction in abundance (Davalos & Esteller, 2010). Further to a simple control of the abundance of microRNAs, the impact on targets can be altered by generation of different isomiRs (Krol et al., 2010). Since the positions of cleavage by Drosha and Dicer determine the 5' and 3' ends of mature microRNA, alternative isomiRs carry alternative seed regions and may target different mRNA. However, it remains unknown to what extent changes in isomiRs affect target binding (Tan et al., 2014). Although one of the strands in the mature microRNA/microRNA* duplex is preferentially loaded into RISC, it has been shown that this selection may change in different organisms, different tissues, and different developmental stages. Arm switching produces microRNA with different targeting properties, which may lead to significant functional consequences (Marco, Macpherson, Ronshaugen, & Griffiths-Jones, 2012). The editing of adenosine to inosine in double-stranded RNA by ADAR proteins can alter microRNA processing if it occurs in pri-microRNA, or alter targeting if it occurs in seed regions of the microRNA. A study using RNA-seq (high-throughput RNA sequencing) in human brain tissue found a clear signal of adenosine to inosine (A-I) editing, and showed that editing generated microRNA with new targeting properties (Alon et al., 2012). Almost 7% of known human microRNAs are in imprinted regions where only either the paternal- or maternal-derived allele is expressed (Girardot, Cavaillé, & Feil, 2012). These imprinted loci have been observed to be dysregulated in schizophrenia, while their status in autism is not known (Gardiner et al., 2012).

1.4.4. MicroRNA conservation

MicroRNAs are present across both animals and plants, with some differences in biogenesis in plants where cleavage by Dicer-like enzymes is performed within the nucleus (Axtell, Westholm, & Lai, 2011). The evolutionary conservation of microRNA sequences reflects the biogenesis pathways, in particular with conservation of cleavage sites (Shi, Gao, & Wang, 2012). Mature microRNA regions are also highly conserved, with seed regions within these particularly conserved (McCreight, Schneider, Wilburn, & Swanson, 2017). The underlying evolutionary pressures driving such conservation are reflected by co-evolution patterns of seed regions and target sites (Barbash, Shifman, & Soreq, 2014).

The expression patterns of microRNAs between cell types and stages vary considerably, with about a third of microRNAs in humans showing high tissue specificity (Landgraf et al., 2007). For example, whereas miR-16 was found consistently abundantly expressed across samples in humans and mouse, miR-21 expression was highly variable (Landgraf et al., 2007). The expression of microRNAs through early development is similarly characterised by conserved patterns across stages (Ninova, Ronshaugen, & Griffiths-jones, 2014). However, it remains unclear to what extent potential microRNA biomarker expression profiles vary by human age.

1.5. MicroRNAs in disease

The study of microRNA in disease is largely dominated by cancer and prospective use of microRNAs as biomarkers. Global reduction in microRNA expression is commonly observed in cancers (Davalos & Esteller, 2010). The use of microRNA in treatment of neurological disease is also under investigation, based on the sensitivity of phenotypes to

microRNA modulation in animal models (Chan & Kocerha, 2012). This could be done through viral and anti-sense-mediated targeting delivered into brain tissue, or nanoparticle delivery of microRNA. MicroRNA biomarkers for disease have been pursued because of the valuable information on the general state of cells that they provide. Although there has not been successful clinical application to date, tests with improved sensitivity and specificity have been developed, for example for Alzheimer's disease (Leidinger et al., 2013). The development of blood tests for early detection of cancer or neurodevelopmental disorders would be a significant contribution to treatment, although better understanding of disease heterogeneity, including subclassification, may be necessary before this is possible. This may also allow personalised medicine in future.

1.5.1. MicroRNAs in brain function

MicroRNAs play an important role in the normal functioning of neurons and specific microRNAs have been found to be especially abundant in the brain (Saba & Schratt, 2010). For example, a study in zebrafish observed defects in neuronal cell differentiation and morphogenesis after *Dicer* was knocked out (Giraldez et al., 2005). MiR-124 (from the three loci mir-124-1, -2 and -3 in humans) has been shown to be key to neuronal cell development. Its overexpression in P19 cells lead to outgrowth of neurites and it is observed to prevent cells from changing towards non-neural transcriptomes. Possible mechanisms include causing neural-pattern splicing by targeting *PTBP1*; knocking down anti-neuronal transcriptional machinery; regulating *BAF* to control histone modifications; down-regulating *SOX9*; and modulation of attachment to the basal membrane through *LAMC1* and *ITGB1* (Visvanathan, Lee, Lee, Lee, & Lee, 2007). Other loci deserving attention include the microRNA in the cluster miR-379-410, which are brain-enriched, and required for activity-dependent dendritic outgrowth of hippocampal neurons (Fiore et al., 2009). Dendritic spines are key sites in neurons (Figure 1.1), where synaptic plasticity allows higher brain functions, by changes in shape, size and number. For this, regulation of

protein synthesis near or within dendritic spines is important. MicroRNA and the silencing machinery have been found at these sites, suggesting that microRNA play a role. For example, miR-134 was found at the synapses of cultured hippocampal neurons, repressing translation of LIMK1 (Schratt et al., 2006). Similarly miR-138 down-regulated dendritic spine size. How the pre-microRNA get to the dendritic spines is an open question – a study showed interesting gradients of abundance for various microRNAs in rat hippocampal neurons (Kye et al., 2007). The microRNA miR-292-5p, miR-26a, miR-26b and miR-25 were particularly well represented at neurite fractions compared to the cell body. Further studies of microRNA interaction with dendritic spine proteins showed that miR-125b and miR-132 are associated with the fragile X protein FMRP (Jin et al., 2004). The latter was also found involved in activity-dependent regulation of neuronal morphology. Together these findings amount to strong evidence of an important role for microRNA in the brain, although they are still restricted to anecdotes about particular interactions and hints at wider significance.

1.5.2. Neurodevelopmental disorders

The study of microRNA in neurodevelopmental disorders is a particularly interesting and growing research field. In schizophrenia, differential expression in cortical grey matter was studied by microarray, revealing dysregulation of let-7g and miR-181b (Beveridge et al., 2008). Pathway analysis showed that the latter commonly targeted genes involved in neural development, as well as several other functions with less obvious interpretations. The targetting of important genes was further supported by *in vitro* experiments. More recently, large GWAS found the strongest association genome-wide with schizophrenia to be produced by a novel variant in miR-137, which appears to target several schizophrenia risk genes as well as nervous system pathways generally (Wright, Turner, Calhoun, & Perrone-Bizzozero, 2013). Following an alternative approach of targetting only SNPs in brain-expressed microRNA for association analysis, a Scandinavian study found a variant in miR-198 to be significantly

associated with schizophrenia (Hansen et al., 2007). Even where microRNA may not be causal themselves, association of variants or differential expression in disease can help to locate their roles in cellular pathways. For example, miR-133b was found to be specifically expressed in midbrain dopaminergic neurons and depressed in Parkinson's disease. Very involved functional studies, including creation of knockout mice, showed that miR-133b was part of a regulatory negative feedback circuit together with *Pitx3* (J. Kim et al., 2007). Regulation by microRNA has been implicated in Parkinson's disease, where variation in a 3' UTR binding site was associated with elevated risk (G. Wang et al., 2008). Investigating the association of variants within a known risk gene, the strongest hit was in a binding site for miR-433, and this allowed a straightforward biological explanation for increased risk. It is thought that similar variants may be responsible for many complex disease associations. Indeed, another study of a candidate risk gene for Tourette's syndrome found a very similar case of variants in a 3' UTR binding site in cases (Abelson et al., 2005). MicroRNA may also be a missing piece in explaining how loss of FMRP causes fragile X syndrome: *in vivo* interaction between FMRP and microRNA, Dicer and AGO1 and loss of function when *AGO1* is knocked out, suggest that its action is mediated by microRNA (Jin et al., 2004). This result is particularly interesting since autism is often comorbid with fragile X syndrome. MiR-132 deserves particular attention as a microRNA very commonly differentially expressed in both psychiatric and neurodegenerative disease, which is also known to function in neuronal plasticity (Chan & Kocerha, 2012).

1.5.3. MicroRNA editing

Adenosine-to-inosine editing of microRNAs in brain tissue was investigated by RNA-seq previously (Alon et al., 2012). Such editing is known to sometimes affect brain function when in protein-coding genes, but there is also a known example of editing affecting targetting for the mouse miR-376. The inosine (I) in the RNA leads to a guanosine (G) in sequencing output, as

well as being read as such by the translational machinery. The human microRNA editing sites that were found had contexts that matched known motifs for A-I editing, and there was conservation with mouse editing sites (Alon et al., 2012). Editing levels at sites were low (<5% of reads aligned to an edit site) but many were in seed regions of microRNAs, and therefore predicted to alter targetting properties.

1.6. MicroRNAs in autism

The first major study of microRNA in autistic brains was conducted using reverse transcription quantitative real-time polymerase chain reaction (qRT-PCR) for 466 targeted microRNAs (Abu-Elneel et al., 2008). The study used actual post-mortem tissue from the cerebellar cortex, with 13 cases and 13 control samples. The cerebellum is one of the brain areas with observed differences in autism at the tissue level, in addition to the inferior olive and the limbic system. Of the originally targetted 466 microRNAs, 227 passed quality controls. Rather than the now-standard differential expression analysis approach, the authors searched for microRNAs in individual autistic samples significantly different to control expression. Based on this, the authors reported 28 dysregulated microRNAs. This approach was not rigorous, however, since the standard deviation of the expression in control samples was estimated, not known (Buyske, 2009). Correction for this yielded only 5 dysregulated microRNAs in two autistic individuals (Buyske, 2009). These statistical difficulties highlight the serious problems in obtaining large enough samples for sensible analysis in such research.

One solution to sample size difficulties is the use of more easily obtained peripheral tissue proxies. This approach was supported in principle by a small study on lymphoblastoid cell lines (LCL) taken from autistic people (Talebizadeh, Butler, & Theodoro, 2008). A microarray assay was performed on RNA samples from which small RNAs were isolated. In the statistical analysis, females and males were considered separately, yielding 9 significantly differentially expressed microRNAs in females (miR-23a,

miR-23b, miR-132, miR-146a, miR-146b, miR-663, miR-92, miR-320 and miR-363). Comparisons between male individuals were then used to confirm a subset of the differentially expressed microRNAs. Autism-associated genes were reported as more frequently targetted by these microRNAs than was the case for randomly selected microRNA – although only one round of random sampling was performed. Following this strategy of expression profiling in lymphoblasts, another study used 14 samples whose mRNA expression profiles had previously been assayed (Sarachana, Zhou, Chen, Manji, & Hu, 2010), which revealed the correlation of target mRNA with microRNA expression. MicroRNA expression was measured using microarray, and Pavlidis template matching was used to find differential expression, which in this case was equivalent to a simple t-test since only two conditions were compared. This revealed 43 differentially expressed microRNA, but it is unclear if any multiple testing correction was applied. Perhaps unsurprisingly, using these 43 probes with principal component analysis and a trained support vector machine both produced easy separation of cases and controls within the training set. Many of the microRNAs had target gene sets which were shown to be associated with functions of embryonic development and neurological diseases by Ingenuity Pathway Analysis. Another study again used microarray to look at microRNA expression as well as mRNA expression levels for 20 affected and 22 control LCL samples (Ghahramani Seno et al., 2011). This found only two differentially expressed microRNA (miR-199b-5p and miR-548o), although a larger number of microRNAs showed large fold-changes in specific sample comparisons. As noted previously, small samples and heterogeneity cause serious problems in statistical analysis, and make collation of results between studies difficult.

Further investigating the potential use of blood samples for characterising microRNA expression in autism, two studies used microarray and validation tests to confirm differential expression of 44 microRNAs (F. Huang et al., 2015) and 1 microRNA (Popov, Madjirova, & Minkov, 2012) respectively. Interestingly, different populations do not appear to produce any more divergent sets of dysregulated microRNAs than for studies in the same

population (F. Huang et al., 2015). Exploration of the networks targeted by the dysregulated microRNAs was a key part of these and other such studies. However, the methods used relied on mapping microRNAs to computationally predicted targets, and then performing enrichment analysis for the target gene set. The total number of enriched pathways when following such a method is usually very large, and often not reported, as in this case (F. Huang et al., 2015). Furthermore, most of the enriched categories appear to have little to do with the condition examined, such as protein digestion and absorption, insulin signalling pathway and many more here (F. Huang et al., 2015).

Another study of microRNAs in serum by qPCR for 55 Japanese autistic individuals and 55 controls similarly found 13 differentially expressed microRNAs, some of which overlapped previous findings (Mundalil Vasu et al., 2014). This study also used the set of differentially expressed microRNAs to test the ability to predict autist status. A similar approach was also used by a study that found 14 differentially expressed microRNAs in autism through RNA-seq (Hicks, Ignacio, Gentile, & Middleton, 2016). The set of microRNAs used to train the logistic regression model was again selected based on differential expression across all samples. This is a flawed approach, because the training set and test set are not truly separated. Although 95% accuracy for autism prediction was reported, the test set on which this was measured was also used to determine the 14 microRNAs used in the model (Hicks et al., 2016).

The use of LCL or post-mortem brain tissue as a proxy for in vivo brain expression has been questioned based on uncertainty in how well peripheral blood samples can represent brain expression, and possible changes in expression after death. As an alternative, olfactory mucosal stem cells were obtained for a study from 8 autistic individuals and 6 controls (Nguyen et al., 2016). Through microarray and qPCR validation, 4 microRNAs were identified as differentially expressed. The signature of dysregulation for these microRNAs was confirmed in skin fibroblasts for different autistic individuals. Further analysis confirmed some specific target predictions for

these microRNAs through luciferase reporter assays (Nguyen et al., 2016). This work shows mixed evidence for and against use of peripheral tissues as biomarkers in autism, and demonstrates the necessity of follow-up qRT-PCR for validation of differential expression.

CNVs found in autism are enriched for microRNA, as shown by a computational analysis of published data (Vaishnavi, Manikandan, Tiwary, & Munirajan, 2013). 11% of autism-associated CNVs previously identified contained microRNA, some of which were previously reported to be associated with autism themselves. Generating a graph of the microRNA edged to the subset of their targets which had previously been implicated in autism, the authors noted some pathway enrichments and hubs of the network, with *DICER1* notably often targetted. These conclusions are weakened, however, by the biased analysis of only the graph with previously implicated targets.

Three further studies have investigated microRNA expression in autism in postmortem brain tissue. These used RNA-seq for 12 autism samples (Mor, Nardone, Sams, & Elliott, 2015), microarray for 10 autism samples (Ander, Barger, Stamova, Sharp, & Schumann, 2015), and RNA-seq for 28 autism samples (Y. E. Wu, Parikshak, Belgard, & Geschwind, 2016) respectively. These reported different microRNAs as differentially expressed in different brain regions. However, it is unclear whether this reflects true differences between regions, or more generally discordant results between the studies outlined here. As autism is highly heterogeneous and outliers with varied causes are common, it is possible that extreme expression of microRNAs in a subset of samples may drive some of the unreplicated findings. This may justify using an approach to differential expression analysis that trims normalised expression to the mean for outlier samples. Cook's distance, a measure of the effect of removing a sample on regression estimates, can be used to achieve this (Anders & Huber, 2010). These studies of microRNA differential expression in the brain identified a total of 84 differentially expressed microRNAs (Ander et al., 2015; Mor et al., 2015; Y. E. Wu et al., 2016). These studies explored possible downstream effects of the

dysregulation through luciferase reporter assays and examination of DNA methylation (Mor et al., 2015). Further exploring of implicated pathways and possible mRNA interactions suggest several good candidates for understanding roles of dysregulated microRNAs in this study (Ander et al., 2015; Y. E. Wu et al., 2016).

The results of studies of microRNA in autism show that there is clearly differential expression in cases, and suggest that differentially expressed microRNA may target genes previously identified in autism. Interesting candidate dysregulated microRNA were highlighted, which need confirmation and further investigation. Studies on LCL demonstrate that peripheral tissue can function as an acceptable proxy for brain expression levels, although shared changes may reflect systemic dysregulation rather than a specific perturbation in the brain. This will allow increased sample sizes, since blood is more easily obtained than brain tissue, and may allow more powerful statistical analyses of differences between cases and controls. It is also notable that most of the studies performed to date have used microarray or qRT-PCR to quantify microRNA expression, approaches that are limited to assaying previously selected microRNA only, and which do not supply the richness of information available from transcriptome sequencing.

1.7. Sequencing methods

Although not providing information on the transcriptome, whole-genome sequencing (WGS) has the advantage of providing information on regulatory regions and a broader picture of genomic variation. For the study of autism, this is important because heritability and additive SNP burden studies suggest many of the susceptibility loci are outside protein-coding regions (Klei et al., 2012). Additionally, CNVs may be a cause of some autism cases, or affect risk more generally, and WGS makes detection of CNVs possible through read depth analysis (Merikangas et al., 2014). Algorithms inspecting spanning reads at a break site also allow more precise

identification of CNV positions.

Several studies have used whole genome sequencing or exome sequencing to survey variation in autism. Discovery strategies using filtering steps for rare and damaging variants, and then convergence on pathways have been proposed (Cirulli & Goldstein, 2010). Rare and *de novo* variants have been identified in autism (L. Liu et al., 2013; Michaelson et al., 2012). However, the nature of rare variation means it is difficult to identify genes enriched for variants as novel candidates (L. Liu et al., 2013). The MSSNG Project is an initiative to build a very large collection of whole genome sequence datasets for families with an autistic member (Yuen et al., 2015). Both parents, as well as probands and some siblings have been sequenced, producing to date 2,300 genome sequences. The findings reported so far show that *de novo* mutations predicted as damaging are more common in autism probands. These *de novo* mutations also appear at CNV breakpoints often, and 75.6% originate from the father (Yuen et al., 2016). Another major finding to date is that known autism genes or loci, even if found to affect a single sibling, only rarely explained autism in both of a pair of siblings. This suggests that the truly causative variants in the majority of cases are still unknown (Yuen et al., 2015). As the MSSNG dataset contains a very large number of samples and variants, the project proposed a more open approach to bioinformatics, with access through a Google BigQuery interface. In this thesis, we present analysis based on a subset of MSSNG data showing variants in microRNA regions.

1.7.1. Transcriptome sequencing

RNA sequencing is an attractive alternative to whole genome sequencing for identifying variants when these are present in transcribed genes. The technique of RNA-seq is very powerful, because it provides information on transcript structure and abundance simultaneously. Read counts allow accurate quantification of expression levels; it allows SNVs to be discovered, concentrating only on the part of the genome that is actually

transcribed; allele-specific expression can be observed and gene fusion and alternative splicing can be detected. Pilot studies showed advantages in cost, while maintaining good power for discovering single nucleotide variations (SNVs) (Chepelev, Wei, Tang, & Zhao, 2009). RNA-seq has also been shown to give consistent results across different laboratories, with microRNA relative expression also reliable, at least for internal comparison within the class of microRNA (Hoen et al., 2013).

Small RNA selection can be used before an RNA-seq experiment to capture microRNA expression rather than to characterise the whole transcriptome. After extraction of total RNA from a sample, a gel is used to separate RNA molecules by size, from which the bands corresponding to microRNA lengths at 20 to 30 nt are cut (C. Lu, Meyers, & Green, 2007). Polyacrylamide gel electrophoresis can be used with manual cutting of the selected band, or automated methods such as the Pippin Prep system which provide higher throughput but less specificity for size (Lopez et al., 2015). The next stage of the library preparation uses ligation of adapters to the 5'-phosphate and 3'-hydroxyl groups at the ends of the selected RNAs. These ligations are followed by purification stages where gel electrophoresis is used to select appropriately sized products (~60 and ~80 nt respectively) (C. Lu et al., 2007). Reverse transcription PCR is then used to obtain cDNA.

1.7.2. SOLiD sequencing

Various methods have been used for transcriptome sequencing previously. These include high-throughput sequencing by ligation in the SOLiD system by Life Technologies, pyrosequencing, and sequencing by synthesis as used in Illumina platforms such as HiSeq (Metzker, 2010). In this work, we use the SOLiD platform, which has an advantage of inherent error filtering due to an overlapping ligation system. In SOLiD (Figure 1.3), a library of DNA fragments (or mate pairs) is prepared with adaptors attached to both ends of each. Clonal bead populations are then grown by emulsion PCR, where each bead starts with a single different sequence attached which is multiplied to

many copies using primers complementary to the adaptors. Bead enrichment follows, and then beads are deposited onto glass slides. Unlike other methods that use sequencing by synthesis, SOLiD uses sequencing by ligation (Applied Biosystems, 2013). In this, fluorescently labelled probes compete to hybridise to complementary sequences adjacent to the primer on the template. Probes that do not hybridise are washed away, and fluorescence imaging reveals the colour of the ligated probe on each bead. The fluorescent label is then cleaved and probes are again allowed to hybridise, this time adjacent to the last probe. By starting this process at different points on the primer and repeating, every position will be assayed twice.

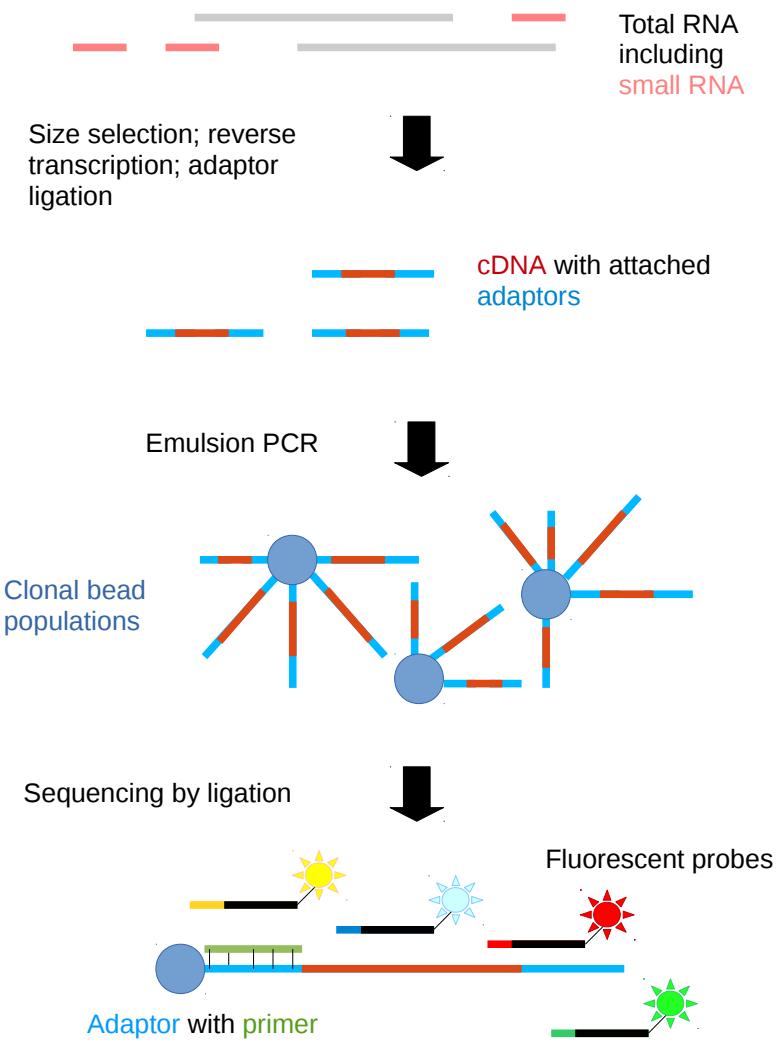


Figure 1.3. The workflow for SOLiD sequencing of small RNA

Extract total RNA from cells of interest, perform size selection to obtain only the fraction of small RNA; perform strand-specific reverse-transcription, ligate 5' and 3' adaptors and run polymerase chain reaction (PCR) with beads; modify bead 3' ends, perform bead enrichment, and deposit onto a glass slide; primers hybridise to adaptors, then repeat cycle of fluorescent probe ligation, detection and cleavage; reset with alternative primer position and repeat.

SOLiD uses an unusual two-base encoding, where there are four different probe colours, each of which is complementary to four different dinucleotide pairs positioned in the middle of the probe (Pekham et al., 2007). In the output sequence, each base is represented by two adjacent colours, and then a SNP would cause both colours to be altered in a sample, whereas a sequencing error would only erroneously change one of the colours. In addition, if two adjacent colours are inconsistent, it is clear that a sequencing error occurred. By filtering reads with such evidence of sequencing errors, accuracy is improved – although in practice SOLiD reads are known to decline in quality severely towards the 3' end (Marco & Griffiths-Jones, 2012).

1.7.3. RNA-seq applications

Because of its virtues, RNA-seq has now been applied to a wide range of biological questions, including discovery of single nucleotide variations (SNVs), differential expression and gene fusion in cancer (Seo et al., 2012), understanding the epigenetic regulatory control of transcription (Dunham et al., 2012) and finding expression quantitative trait loci and variants underlying alternative splicing (Battle et al., 2013; Lappalainen et al., 2013; Montgomery, Lappalainen, Gutierrez-Arcelus, & Dermitzakis, 2011). Recent advances include the development of single-cell sequencing, which has shown that within an apparently homogeneous tissue, single cells may have remarkably different transcriptomes. For example, bimodal expression of immune genes was observed in bone-marrow-derived dendritic cells (Shalek et al., 2013). Relevant to this project is the important field of small RNA characterisation, where RNA-seq is applied to an isolated short-length fraction of total RNA (Ozsolak & Milos, 2011). For example, small RNA sequencing was used to investigate microRNA in switchgrass used for biofuel (Xie et al., 2013). Algorithms were used to identify novel microRNA, to discover differential expression between different growth conditions, to predict microRNA targets and to annotate predicted targets with gene ontology (GO) terms and Kyoto Encyclopedia of Genes and

Genomes (KEGG) molecular functions.

1.7.4. Computational analysis of RNA-seq data

1.7.4.1. Alignment

The high-throughput sequencing approach to genomics is founded on the use of algorithms and software packages to analyse data. At the beginning of an analysis, alignment to a reference genome or transcriptome is required, after which several programs can process the aligned read data (Hong et al., 2012). Alignment depends on indexing of the target genome by k-mer in memory, so that a fast seed-and-grow approach can be used, as in Bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009). In the case of our study, as small RNA sequencing was performed with SOLiD, sequentially removing the end colours from reads until they successfully map to the genome by Bowtie is necessary due to decreased 3' read quality. We therefore use the alignment tool SeqTrimMap (Marco & Griffiths-Jones, 2012). A similar trimming approach to achieve alignment to reference hairpins has also been reported in isomiRID (de Oliveira, Christoff, & Margis, 2013).

1.7.4.2. Variant calling

Given aligned reads, the software SamTools can manipulate alignment files and call variants based on alternative bases aligned at reference genome positions (H. Li et al., 2009). This is performed based on Bayesian inference, using the Phred score of aligned bases to infer the likelihood of a mismatch for the read at the position. More sophisticated pipelines have been developed for local realignment near indels and adjusted quality measures in GATK, however these tools have not been applied previously for SeqTrimMap alignments (Nielsen, Paul, Albrechtsen, & Song, 2011).

1.7.4.3. Differential expression analysis

Differential expression analysis uses read count data for multiple samples to attempt to identify genes that have significant differences in expression between sample groups. In the general case of gene differential expression analysis, alternative splicing and bias due to gene length must be accounted for. Expression levels of isoforms are treated as parameters of a statistical model, and the parameters with the best likelihood are found by convex optimisation (Jiang & Wong, 2009). The case of microRNA is simpler, and 5' isomiR expression can be quantified by exact read start positions, rather than use of FPKM (Finotello & Camillo, 2014). For quantification of microRNA expression and isomiR expression, HTSeq-count and miRSpring have both been used previously, although the latter is not designed for SeqTrimMap alignments (Anders & Huber, 2010; Humphreys & Suter, 2013). Given such gene expression count data for case and control samples, count distributions are modelled as following a negative binomial distribution, whose parameters can be estimated based on the observations (Anders & Huber, 2010). Such models are necessary because, although RNA-seq is known to be reproducible and produce largely accurate rankings of microRNAs by expression level, read counts are not directly proportional to expression (Hoen et al., 2013). Some genes have more variable expression, and read counts for genes can be too low to determine this variability for each gene separately. Therefore, DESeq works by estimating variability as a function of mean expression (Anders & Huber, 2010). Genes are then called significantly differentially expressed based on the difference in normalised expression between sample groups.

Due to the very large number of microRNAs assessed for differential expression, multiple testing correction is necessary. Benjamini-Hochberg adjustment of raw p-values can provide controlled false-discovery rates (Benjamini & Hochberg, 1995). In this algorithm, tested items are ranked ascending with their p-value from item-by-item hypothesis tests, and accepted as 'discoveries' only until a p-value is encountered greater than the

item rank divided by remaining items plus one, multiplied by desired false discovery rate (FDR). The FDR can then be interpreted as the broad probability that the 'discoveries' are true. In this work, we refer to FDR unless otherwise specified when multiple testing correction is mentioned.

The alignment of sequencing reads to multiple locations in the genome also presents problems for differential expression analysis. This is particularly important in the case of microRNAs, where similar or identical mature products commonly originate from multiple loci. Such reads cannot be discarded without losing information on a large number of microRNAs, and strategies for RNA-seq that use read information across larger genes are not applicable (Finotello & Camillo, 2014). While inclusion of reads with multiple alignments is a conservative addition for tests of significant differential expression between conditions, multiple possible source loci must be considered in the analysis of results.

1.7.4.4. Partitioning of sample data

Among the difficulties in RNA-seq analysis for autism are the high variability within autism for phenotypes and outlier cases. Some studies have addressed this issue by computational partitioning before proceeding with analysis of subgroup characteristics (V. W. Hu & Steinberg, 2009). Clustering algorithms were used based on item scores on the ADI-R for approximately 2000 individuals, and three subgroups were identified. A similar investigation of the structure of the ADI-R also found three clusters with factor analysis suggesting two latent factors representing severity of social and communication behaviours and repetitive and restricted behaviours (Snow et al., 2009). Clustering approaches include visual inspection for separation through principal components analysis, k-means clustering and other unsupervised learning methods (Hall et al., 2009).

MicroRNA expression patterns have previously been proposed as effective tools for diagnosis of conditions such as Alzheimer's disease, as well as

autism (Leidinger et al., 2013). To be considered a valid biomarker, proposed microRNAs must be accurately and reproducibly measurable, and provide an objective indication of an individual's state. The reliable quantification of microRNAs by RNA sequencing and qRT-PCR is supported by assessments of variability under those techniques (Hoen et al., 2013). However, variation across human development and the ability of peripheral blood to reliably predict the state in other tissues remains a challenge for microRNA biomarkers (Leidinger et al., 2013).

Biomarkers may be classed as diagnostic (where they indicate the presence of a disease or a condition for a patient), prognostic (where they indicate a patient's likely outcome, such as disease progression), or predictive (where they predict treatment response) (Hamam et al., 2017). For example, gastric cancer recurrence and patient survival may be predicted by hsa-miR-335 expression (Z. Yan et al., 2012). Similarly, stratification of non-small cell lung cancer cases has been proposed using hsa-miR-146b expression (Raponi, Dossey, Jatkoe, Wu, & Chen, 2009). MicroRNAs also have strong potential as diagnostic biomarkers in cancer, with tumor-derived microRNAs existing in stable form in sputum and blood, such as hsa-miR-21 in non-small cell lung cancer (Davalos & Esteller, 2010).

1.8. MicroRNA target prediction

Computational prediction of the targets of microRNAs is used to identify likely regulated genes and pathways. These methods include naive approaches based on seed pairing such as SeedVicious (Marco et al., 2012). Another method uses search for complementary sequences in a set of 3' UTR sequences and then calculate free energy between candidate microRNA-mRNA pairs in miRanda (Enright et al., 2003). Other models include the energy cost of getting access to potential target sites in PITA (Kertesz et al., 2007), machine learning with coding sequence binding allowed in DIANA-microT-CDS (Reczko et al., 2012), and target site conservation in TargetScan (Lewis, Shih, Jones-Rhoades, Bartel, & Burge,

2003). Experimental methods for surveying microRNA binding genome-wide by CLIP-seq and for testing specific interactions through luciferase reporter assays provide less comprehensive, but likely more reliable datasets (Papadopoulos, Reczko, Simossis, Sethupathy, & Hatzigeorgiou, 2009).

1.8.1. Computational discovery of microRNA function

Given a discovery of microRNA involved in a biological condition, several computational approaches are possible to obtain an understanding of role or function. For studies identifying large numbers of genes, such as those affected by CNVs, analysis of networks implicated and clustering (Gilman et al., 2011) or simple enrichment analysis is possible (D. W. Huang, Sherman, & Lempicki, 2009). The primary difficulty in this situation is the separation of microRNA from annotations of biological processes and functions, and the large number of targets for each microRNA (Huntley et al., 2016). A common approach to this problem has been to predict the targets of affected microRNAs, and then proceed with functional enrichment analysis using the targeted genes as a starting point (S. Y. Lee, Sohn, & Kim, 2012). However, this has led to studies which have produced hundreds of affected pathways, biased towards commonly appearing pathways or gene ontology classes (D. W. Huang et al., 2009; T.-P. Lu et al., 2012). In this project, we will present an analysis of bias in microRNA functional enrichment analysis, demonstrating problems with these commonly used approaches. This work has led to other developed software which address the problem of functional enrichment directly through annotation of microRNAs (Backes, Khaleeq, Meese, & Keller, 2016).

1.9. Project aims

1.9.1. General aims

The aim of this PhD project is broadly to investigate the role of microRNA in autism susceptibility. We will attempt to characterise the microRNA profile of autistic individuals, to identify differences with non-autistic individuals, and to explore the participation of microRNA in functional pathways contributing to autism. As described above, there are also some practical concerns about autism research and microRNA analysis in general. In this work, we aim to contribute to improved statistical methods for functional enrichment analysis. We will also address some of the barriers to community engagement and produce recommendations for building bridges for more successful future research.

1.9.2. Investigation targets

Under the hypothesis that microRNA may be an important missing piece to fill gaps in knowledge about autism, we will investigate several major areas:

(1) A portion of the missing genetic heritability for autism may be due to rare (so poorly tagged) variants in microRNA. Such variants could very plausibly cause the wide range of autism phenotypes due to the pleiotropy of microRNA, and strong conservation of microRNA (Chan & Kocerha, 2012) implies that the effects of such variants would be severe. We therefore investigate the presence of rare variants in autistic subjects.

(2) Some studies of microRNA in autism to date have been hampered small sample sizes, and statistical analyses have therefore been limited (Hicks & Middleton, 2016). Furthermore, only a subset of microRNAs were generally assayed. We attempt to address some of these problems with a RNA-seq study identifying differential expression. We further explore the networks

and functions that may be implicated by these changes in expression.

(3) A-I editing in microRNA, arm switching and isomiR differential expression are interesting areas of emerging research. We investigate these features of the transcriptome using small RNA sequencing data to look for unexpected dysregulation.

(4) MicroRNA may have utility in prediction of autism as a peripheral tissue biomarker. A key focus of autism research is improving early diagnosis, so that support and intervention may be applied earlier and more effectively (Klintwall, Eldevik, & Eikeseth, 2013). A consistent microRNA expression signal may be complementary to other methods of early prediction, such as eye contact decline (Jones & Klin, 2013). In this work, we present an attempt to use machine learning approaches to predict autism status through microRNA expression.

(5) We attempt to address problems with computational approaches to identifying microRNA functions in diseases in general (S. Y. Lee et al., 2012). Having identified bias in commonly used methods, we present alternatives and argue for better practice.

(6) We aim to investigate how research into autism could be improved in general, and to find the opinions of autistic people about autism research. We aim to use this to produce proposals for the research community to improve engagement.

2. Investigation of microRNAs in autism by small RNA sequencing

2.1. Abstract

Autism is a highly heritable neurodevelopmental disorder with a wide range of phenotypes, at the core of which are impairments in social communication and interaction and repetitive or restricted behaviour and interests. Studying the genetic basis of risk provides an opportunity to understand the underlying biology of autism, and may improve the implementation of early intervention treatments. Previous studies have shown microRNA dysregulation in autism, and suggested that further investigation may reveal important regulatory roles. Here we report small RNA sequencing of lymphoblastoid cell lines for 42 autistic individuals and 10 matched controls. Using this sequencing data, we have identified differences in expression, variants and microRNA processing. We found 24 microRNAs which were differentially expressed in autism, and 8 microRNAs which had significant switches in arm usage. We performed validation for differential expression through qRT-PCR. The set of dysregulated microRNAs included several microRNAs previously identified in autism, such as miR-23a-3p and miR-146a-5p as well as novel candidates with previous associations to neurodevelopmental disorders. We found shared pathways implicated in dysregulation for several of these microRNAs involving Fragile X protein and dendritic growth.

2.2. Background

Autism is a neurodevelopmental disorder defined by difficulties with social communication and social interaction, and repetitive and restricted behaviours and interests. Understanding the genetic basis of autism risk is important because some of the impact on affected individuals and carers may be mitigated with early interventions (Klintwall et al., 2013). Exome

sequencing projects (L. Liu et al., 2013), copy number variation analyses (Gilman et al., 2011) and genome-wide association studies (Anney et al., 2012) all point towards a complex combination of variants in numerous genes. Among the loci affected, microRNAs are of particular interest because of their important role in the brain and ability to regulate large numbers of genes (Schratt et al., 2006; Somel et al., 2011; B. Xu, Karayiorgou, & Gogos, 2010). Rare variants in microRNAs, which are not effectively tagged in genome-wide association studies, could plausibly cause broad autism phenotypes due to the pleiotropy of microRNAs. Strong conservation of microRNAs also suggests that the effects of such variants would be severe.

A small number of studies have found differential expression of microRNAs in autism using qRT-PCR (Abu-Elneel et al., 2008) and microarray assays (Ghahramani Seno et al., 2011; Sarachana et al., 2010; Talebizadeh et al., 2008). Other investigations have used RNA-seq more recently (Hicks et al., 2016; Mor et al., 2015; Y. E. Wu et al., 2016). The targets of affected microRNAs were reported to be enriched for known autism genes (Talebizadeh et al., 2008), although there may be bias to such analyses (Bleazard et al., 2015). MicroRNAs found differentially expressed in the brain in autism were also replicated when investigating peripheral blood. Studies using only brain tissue suffer difficulties with statistics because of tissue availability leading to low sample sizes, which may be alleviated with the use of blood samples (Buyske, 2009; Talebizadeh et al., 2008). These studies also suggest that microRNA expression levels may have utility as a biomarker predictive of disease state (Sarachana et al., 2010). If a predictive model could be constructed for larger samples with reasonable accuracy, then it could become a powerful tool in identifying children who would benefit from early intensive behavioural intervention (Klintwall et al., 2013).

In this study, we report analysis of small RNA sequencing data for 42 autistic individuals and 10 age-matched individuals without autism diagnosis. We use the sequencing data to discover microRNA expression

levels, while also identifying variants and other transcriptome features such as A-I editing and isomiR expression.

2.3. Methods

2.3.1. Sample details

Lymphoblastoid cell lines were obtained from 42 autistic individuals and 10 age-matched individuals without autism diagnosis (referred to as controls). UK and Dutch individuals with autism were selected from the International Molecular Genetic Study of Autism Consortium (IMGSAC) project (Hervas et al., 1998). These were required to be from multiplex families, due to known differences in the genetic component for multiplex and simplex autism (Klei et al., 2012). We selected only IMGSAC individuals self-reporting as Caucasian, or whose ancestry analysis identified them as such as previously reported (Hervas et al., 1998). Individuals with unknown gender were excluded. Samples were also selected for those which had full genotype data available (although this was not used in this study) and no known autism CNVs reported previously (Pinto et al., 2010). We obtained cell cultures for 10 British individuals without autism diagnosis from the Coriell Collection (Consortium, 2015).

EBV transformed lymphoblastoid cell lines were cultured as described previously (IMGSAC, 2001). Iterative tests were performed to optimise procedures for RNA extraction to maximise microRNA content and yield in subsequent library preparation and sequencing. Total RNA was extracted from $1x10^8$ cells for each sample stored in RNALater using Ambion miRVana extraction kits. The use of human derived biological material in this research was covered by the Oxfordshire Research Ethics Committee A, 003.013.

2.3.2. RNA sequencing

Small RNA libraries were prepared and sequencing was performed using AB/Life Technologies SOLiD by the University of Manchester Faculty of Life Sciences sequencing facility. We annotated samples alternatively with their sequencing batch forming 4 groups. The samples were sequenced at the University of Manchester on 2011/10/13, 2012/02/24, 2012/04/17 and 2012/09/27. Samples were prepared for sequencing on the SOLiD 4.0 analyzer (Life Technologies) according to manufacturer's instructions, with barcoded small RNA libraries prepared using the SOLiD Total RNA Seq kit from Ambion. Library quality was confirmed using the Agilent Bioanalyzer. Libraries were pooled in equimolar ratios. The EZ bead system was used to produce templated beads for pooled library sequencing, and these were deposited and sequenced following standard protocol for the SOLiD 4.0 analyzer.

2.3.3. Sequence alignment and variant calling

Raw sequencing reads were filtered using an in-house bioinformatics script that removed reads failing criteria requiring that the first 10 colours have a minimum of 3 over quality threshold of 22, and that the total colours below quality threshold 9 in the read is below a maximum of 10. The reference human genome hg19/GRCh37 downloaded from UCSC was used to build a Bowtie index ready for alignment (Langmead et al., 2009). Alignment was performed using the SeqTrimMap wrapper for Bowtie with standard parameters (up to 5 multiple alignments allowed and 2 mismatches allowed) (Marco & Griffiths-Jones, 2012). SAMTools v1.19 was used to convert output SAM alignments to sorted BAM files, allowing viewing with IGV and variant calling (H. Li et al., 2009). Bcftools (part of the SAMTools package) was used to call SNVs from the output alignment. A custom script was then used to collect all SNPs identified in miRBase release 20 annotated microRNAs and join this to other information such as seed membership and SNP identifier from dbSNP build 139. SNVs required a

Phred score above the threshold of 17.0 for inclusion.

2.3.4. MicroRNA expression

A script provided with Python package HTSeq-count was used to count the number of reads for each microRNA in each sample using miRBase release 20 annotations (Kozomara & Griffiths-Jones, 2011). DESeq2 was used in R to move from the raw HTSeq counts to measurement of differential expression (Anders & Huber, 2010). Where DESeq2 identified microRNAs which had outliers that were distorting results according to Cook's distance, with threshold calculated by default, these were replaced with trimmed means, as a conservative alternative to exclusion from analysis. Following the standard DESeq2 process, heatmaps and principal components analysis (PCA) plots of the data were also produced, along with tests for differential expression.

2.3.5. Scripts for read analysis

We investigated several details of expressed microRNA using custom scripts with HTSeq modules used to read BAM format sequencing data. As an alternative to simple counting of reads overlapping annotated microRNAs, a script was used which counted the number of reads supporting each isomiR that was in an annotated pri-microRNA. Due to the nature of our alignment process trimming 3' bases, we only attempted to identify processing of mature microRNAs with alternative 5' start sites, which we called '5' isomirs'. All reads supporting each 5' isomiR were counted and differential expression analysis was performed with this data. This used the same pathway for differential expression analysis described above, but with modified input data with entries for transcripts for each 5' isomiR. We used Mann-Whitney U-tests to investigate whether the fraction of reads from the dominant isomiR was different in autism for each primary microRNA. Using miRBase annotations, we matched pairs of 5p and 3p arms for each primary microRNA, and counted reads supporting each arm. We then

performed Mann-Whitney U-tests for each pair to identify significant changes in arm processing between autism and controls.

We used read data to investigate A-I editing by counting the number of reads supporting reference adenine against the number supporting guanosine (by which inosine is read in our sequencing experiments). Due to the very large number of sites with reference A in our sequence data, we first assessed a set of sites previously identified with strong evidence for A-I editing in human brains (Alon et al., 2012). We used a Phred score threshold of 30 for base calls at those sites to select and count reads supporting G. Using the principle that all accepted base calls had an estimated error rate of less than 0.001, we tested with the binomial distribution to make a conservative assessment of significantly high G read counts. We calculated the binomial distribution of the number of non-reference reads aligned at a position under the null hypothesis of no variant at the location, using a probability for each aligned read of 0.001 for sequencing error. The observed count of reads supporting G at a position was then tested against this distribution, giving a one-sided p-value for this null hypothesis.

2.3.6. Phenotype assessment

For an analysis of autistic individual phenotype, we used a wider dataset including all autistic subjects in our study, as well as others who participated in the IMGSAC project (IMGSAC, 2001). Written informed consent for phenotype and genotype data collection was obtained previously from parents and guardians and where possible autistic individuals. Initial screening questionnaires were completed by parents, and these were then followed by the Autism Diagnostic Interview-Revised (ADI-R). The ADI-R contains 37 central items based on direct observation score and other questionnaire answers given by parents during the interview. Previous diagnosis of these individuals was based on combining ADI-R scores following the test specifications. Under the IMGSAC project, further assessment was also carried out using the Autism Diagnostic Observation

Schedule (ADOS), which was used to confirm ADI-R assessments, and also combined with Vineland Adaptive Behaviour Scales scores and other psychometric tests not used here.

2.3.7. Variant validation

In order to investigate the validity of rare seed region variants identified from RNA-seq data, we performed polymerase chain reaction (PCR) to amplify genomic DNA for selected regions, which were then sequenced by the DNA Sequencing Facility at the University of Manchester by Sanger sequencing. Genomic DNA for selected samples was obtained previously as part of the IMGSAC project by Nucleon kit from blood samples prior to the establishment of lymphoblastoid cell lines (Hervas et al., 1998). Primers were designed using Primer3 with standard settings and confirmed as unique to the selected region using Repeatmasker and in-silico PCR through the UCSC Genome Browser. We inspected the generated chromatograms using the EMBOSS program Abiview (Rice, Longden, & Bleasby, 2000).

2.3.8. qRT-PCR

We used quantitative real-time PCR experiments to validate differential expression signals. We selected three microRNA endogenous controls for normalisation, based on manufacturer recommendations, as well as confirmation through RNA-seq data of consistent expression ($p > 0.5$ differential expression in autism samples, normalised range within 0.25 times expression mean). We normalise within and confirm repeatability within plates (Bustin et al., 2009). Each control and test microRNA for each sample had three technical replicates within plates, in addition to negative controls. RNA samples were prepared as described above, using Ambion miRVana extraction kits. RNA sample quality was confirmed using Agilent Bioanalyzer for 12 samples with RNA Integrity Number 9.8-10. We used Taqman Advanced microRNA Assays designed to target test microRNAs designed by the manufacturer. Taqman Advanced microRNA cDNA

Synthesis Kits were used following standard procedures for poly(A) tailing, ligation, reverse transcription and miR-Amp reactions. Final qRT-PCR amplification was performed using the Taqman Fast Advanced Master Mix with the Applied Biosystems Quantstudio 12K Flex using 384-well comparative Ct fast reaction tests. We obtained amplification plots and Ct values from the manufacturer automatic Ct calculation algorithm. We required that three replicates for each sample microRNA pair were all successfully amplified with no amplification in negative controls, standard deviation between replicates below 0.3, and no fails due to other Quantstudio QC filters (bad passive reference signal, baseline algorithm fail, high noise, replicate outlier, or noise spikes). We removed sample microRNA pairs for which these requirements failed, and tested for differential expression with only passing pairs. Normalisation was performed by subtracting the average Ct score for three endogenous control microRNAs for each sample. We then performed a standard T-test for different delta Ct scores for each microRNA in passing autism and control samples (Schmittgen & Livak, 2008).

2.3.9. Machine learning

The machine learning utility *scikit-learn* v16 for Python was used to train a classifier for autism status based on normalised microRNA expression profiles and tested for accuracy with leave-one-out cross-validation (Hall et al., 2009). We used raw read count data for each mature microRNA for each sample, which was then processed to scale the microRNAs as features to have zero mean and unit variance. We used a linear kernel support vector machine with a regularising factor of 1 as the penalty parameter for error as default. We performed logistic regression with L1 norm regularisation and a regularising factor of 1. We also applied a random forest classifier. We performed feature selection to identify potential biomarkers using our support vector machine classifier with stronger regularisation (c=0.05) in order to find more consistent classifying microRNAs. We explored correlation in expression levels among microRNAs and inspected

correlations with differentially expressed microRNAs and selected features using Python scripts for normalised count data.

2.3.10. Network analysis

We used the empirical functional enrichment analysis tool which we developed (described later in this thesis) to investigate over-represented gene ontology categories for our differentiall expressed microRNAs (Bleazard et al., 2015). We used miRanda version 3.3a requiring free energy under 20kcal per mol and target score above 155 with the target database Ensembl release 75 and a basic and multihit model with known autism genes from the literature included as a tested category (Basu et al., 2009). We also used the analysis tool miEAA for over-representation analysis using target predictions by the miRWalk algorithm and testing for classes with at least two targeted members (Backes et al., 2016). We explored connections between differentially expressed microRNAs and direct and experimentally confirmed gene interactions using Ingenuity Pathway Analysis (IPA). We used Knowledgebase connections between targeted genes and IPA visualisation to explore possible connections to candidate autism networks.

2.4. Results

2.4.1. Sequencing and alignment

A total of 692,123,038 raw reads were generated after filtering from the 52 samples. This produced 973,663,579 alignments to the human genome from 617,621,255 reads after trimming with SeqTrimMap (Figure 2.1). Reads were allowed up to 5 alignments with up to 2 base mismatches. The overall proportion of reads which mapped to the genome was 89%, suggesting reasonable general data quality. One sample (JL_18) had a particularly low proportion of mapped reads at 39%, while no others were below 69%. We measure the coverage for microRNAs achieved by sequencing by the

median number of reads for all annotated microRNAs across all samples. However, as a large number of annotated microRNAs are not expected to be expressed at detectable levels within our samples, we exclude from this assessment those with zero aligned reads. We find the median number of reads aligned to a microRNA in a single sample to be 4,557.

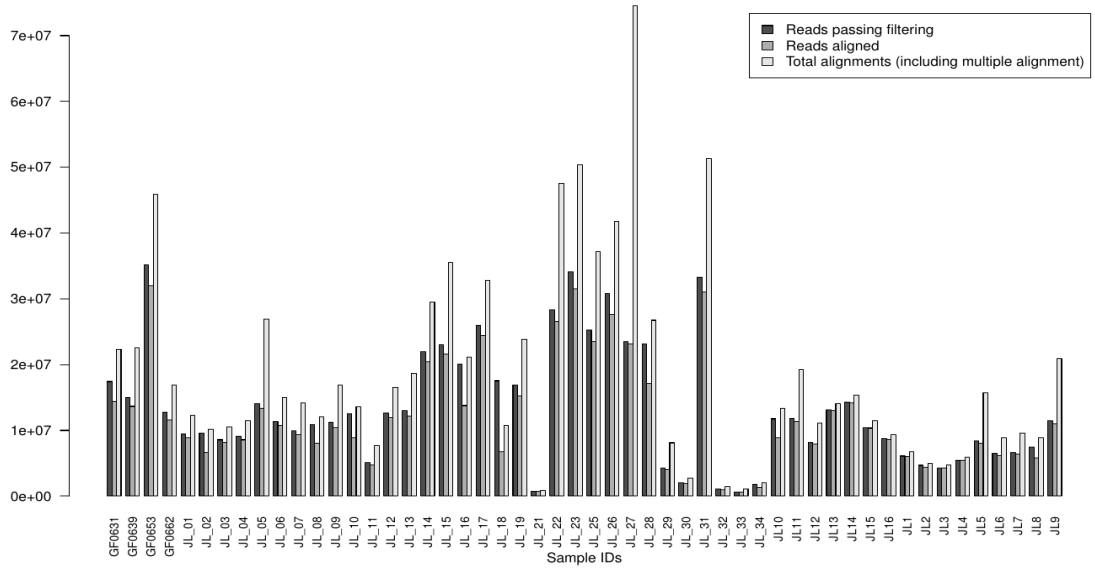


Figure 2.1. Read counts and alignments for sequenced samples

Counts of total reads passing filtering for each sample, total aligned reads, and total number of alignments (including multiple alignments) to the genome.

2.4.2. Variant detection

In order to identify sequence variants within microRNAs that were present in autistic individuals and may be involved in genetic risk, we followed a variant calling pathway. Bcftools, from SAMTools v1.19 was used to call variants for the output alignments using default parameters (H. Li et al., 2009). A total of 499,566 variants were called among all 52 samples. We used a script to select those variant calls which passed a threshold for Phred quality of 17.0, and which were located within an annotated microRNA. We selected this threshold as one that gives a probability of error for calls of 0.02. We used miRBase v20 to map variants to annotated microRNAs, using records for 1,871 pre-microRNAs and 2,794 mature microRNA region records. We also matched the SNPs which were present in dbSNP with their annotations there and identified variants present in microRNA seed regions. We used the dbSNP set of 'common' SNPs, which refers to those which are present at about 1% frequency in at least one dbSNP population (Sherry et al., 2001). Only 47 variant positions passed these criteria, 7 of which were found in seed regions. Of the 47 passing variants, 22 were outside of annotated mature microRNAs, instead identified from reads aligned to pre-microRNA regions. We investigated further the remaining 25 variants within mature microRNAs (Table 2.1). We also tested for variants remaining under alternative stricter criteria requiring a Phred quality score above 30.0, giving probabiliy of error of 0.001 for such variant calls. Of the original 47 passing variants, 28 also passed the stricter criteria, none of which were present in seed regions.

Table 2.1. List of microRNA variants

Chromosome	Position	Variant Allele	Total Autism Samples	Common dbSNP Record	Mature miRNA	In Seed Region
1	17375	G	1		hsa-miR-6859-3p	
1	567783	A	1		hsa-miR-6723-5p	
2	56210164	C	1		hsa-miR-217	
2	177015068	C	1		hsa-miR-10b-5p	
2	207647981	T	1	rs2241347	hsa-miR-3130-5p	
2	207648009	A	2	rs115772313	hsa-miR-3130-3p	
3	15537801	G	6		hsa-miR-4270	True
3	186504489	C	1		hsa-miR-1248	
5	71465308	C	0		hsa-miR-4803	True
5	72174432	G	0	rs266435	hsa-miR-4804-5p	True
5	159912418	G	8	rs2910164	hsa-miR-146a-3p	True
6	31323020	C	1		hsa-miR-6891-3p	True
7	5535483	C	1		hsa-miR-589-3p	True
9	97847805	C	1		hsa-miR-27b-3p	
11	93466866	T	35	rs2155248	hsa-miR-1304-3p	
11	96074619	G	1		hsa-miR-1260b	
11	113320793	C	1		hsa-miR-4301	True
12	17826251	C	1		hsa-miR-3974	
12	17826252	A	1		hsa-miR-3974	
12	49165801	C	1		hsa-miR-4701-5p	
12	121882098	G	1	rs3817551	hsa-miR-7107-3p	
15	102513788	C	1		hsa-miR-	

Chromosome	Position	Variant Allele	Total Autism Samples	Common dbSNP Record	Mature miRNA	In Seed Region
					6859-3p	
16	67058	G	1		hsa-miR-6859-3p	
21	17912169	A	2		hsa-let-7c-5p	
22	38243748	T	1		hsa-miR-659-5p	

List of SNVs passing variant quality thresholds with Phred score above 17.0 identified within mature microRNA regions. Counts of the number of autism samples out of the total 42 with the SNV called are shown. The dbSNP rs-numbers for SNPs are shown where they are present in the database of common variants. SNVs are marked as present in the seed region of a microRNA where they are between bases 2 and 8 from the start of the mature microRNA.

2.4.3. Sanger sequencing for variant validation

Since SNVs called by our pipeline were sensitive to quality thresholds, we attempted to validate them through Sanger sequencing of genomic DNA. In particular, seed regions of microRNAs are believed to be the most important for determining targeting properties, and microRNAs are highly conserved in these regions (Enright et al., 2003). As 5 out of the 7 SNV calls in these regions were not at common dbSNP loci, there is a high prior probability of false positives for such calls. We attempted the validation of 3 SNVs passing detection criteria with Phred scores above 17.0 and present in microRNA seed regions using Sanger sequencing. The remaining 4 SNVs called in seed regions were not assessed due to multiple controls called for the SNV (hsa-miR-4270), not appearing in an autistic subject (hsa-miR-4803), and appearing as a common SNP in dbSNP (rs266435 and rs2910164). We tested five other candidate variants which were not found in seed regions but were otherwise of interest (hsa-miR-3974, hsa-miR-3130-3p, hsa-miR-217, hsa-let-7c-5p and hsa-miR-659-5p). These were selected for several reasons, including let-7c-5p which was identified as differentially expressed. Small RNA sequencing reads supporting variants to be validated were inspected using IGV. Primers for amplification were designed using primer3 (Table 2.2). To confirm unique mapping to the human genome for the primers, we performed alignment with Repeatmasker and in-silico PCR through the UCSC Genome Browser. IDs for LCL samples were matched to stored genomic DNA samples, and also mapped to family members.

Table 2.2. List of PCR primers for validation

Chromosome	Position	miRNA	Forward Primer	Reverse Primer	Samples Tested
2	56210159	hsa-miR-217	GCATCTTGGGCTCACCTCTC	ATCTCTCTTGGTAAATTGGGAAAATT	3
2	207648008	hsa-miR-3130-3p	TGAGTCAAATGCCCATCCCT	TGAAGCCAGGGGAGTAAAGG	1
6	31323020	hsa-miR-6891-3p	GGAACACTTCTACCTGGGGC	GAGCTTGTGGAGACCAGACC	1
7	5535483	hsa-miR-589-3p	CACCAAAGAGACTCCGGCC	GAGCAGGGCCATGGAGAAG	1
11	113320793	hsa-mir-4301	AGAAAGAAATCAGGGGGCGG	AGCTGGAAGTGCTTTGGCTC	1
12	17826252	hsa-mir-3974	TGTTTGGGTCTGTGCTTCTGAT	TGTTTAAACAAGGAGTGGGCAAT	9
21	17912169	hsa-let-7c-5p	GTGACCTATGCTGGAAACCCA	AGGGCTGAAATGATCATTTTAATGT	2
22	38243748	hsa-miR-659-5p	TGGAAGATGAGGGTTGGTTAGA	CCTTCCCCATCGGTATCTGC	1

Locations of SNV calls selected for validation. Primer sequences were designed using Primer3.

After amplifying genomic regions and Sanger sequencing, we inspected chromatograms using the EMBOSS program Abiview (Rice et al., 2000). Amplification failed for two of the samples. We compared the observed sequence to human genome reference for the remaining 15 tests, of which 14 matched the human reference genome sequence, suggesting that the RNA-seq variant calls were false positives (Table 2.3).

Table 2.3. Sanger sequencing results

microRNA	Test Outcome	Variant Position Offset on Read
hsa-let-7c-5p	Human genome reference sequence	183
hsa-miR-217	Human genome reference sequence	150
hsa-miR-217	Human genome reference sequence	151
hsa-miR-3130-3p	Sequencing failure	N/A
hsa-mir-3974	Human genome reference sequence	186
hsa-mir-3974	Human genome reference sequence	188
hsa-mir-3974	Human genome reference sequence	190
hsa-mir-3974	Human genome reference sequence	187
hsa-mir-3974	Human genome reference sequence	189
hsa-mir-3974	Human genome reference sequence	189
hsa-mir-3974	Human genome reference sequence	186
hsa-mir-3974	Human genome reference sequence	187
hsa-mir-3974	Human genome reference sequence	186
hsa-mir-4301	Human genome reference sequence	177
hsa-miR-589-3p	Human genome reference sequence	194
hsa-miR-659-5p	Variant confirmed	160
hsa-miR-6891-3p	Sequencing failure	184

Tests for combinations of candidate microRNA variants with samples in which those variants were called from RNA-seq data. 14 of the validation tests showed that called variants were likely false positives, as sequencing product matched the human reference genome. Evidence for a variant was only confirmed in hsa-miR-659-5p. The variant position offset on read refers to the position on the Sanger sequencing product to which the putative variant aligned.

We confirmed one variant on hsa-miR-659-5p in the seed region (Figure 2.2). This variant was previously observed in small frequencies in HapMap reference populations, with reference rs369207831. Interestingly, hsa-miR-659-5p target sites were previously reported to be affected by common mutation in the gene GRN, which is associated with neurodegenerative disease (Rademakers et al., 2008). This variant was present in only a single autistic individual. The other tested variants failed validation. We tested for but did not find evidence for misalignment in the RNA-seq data at these locations using BLAT.

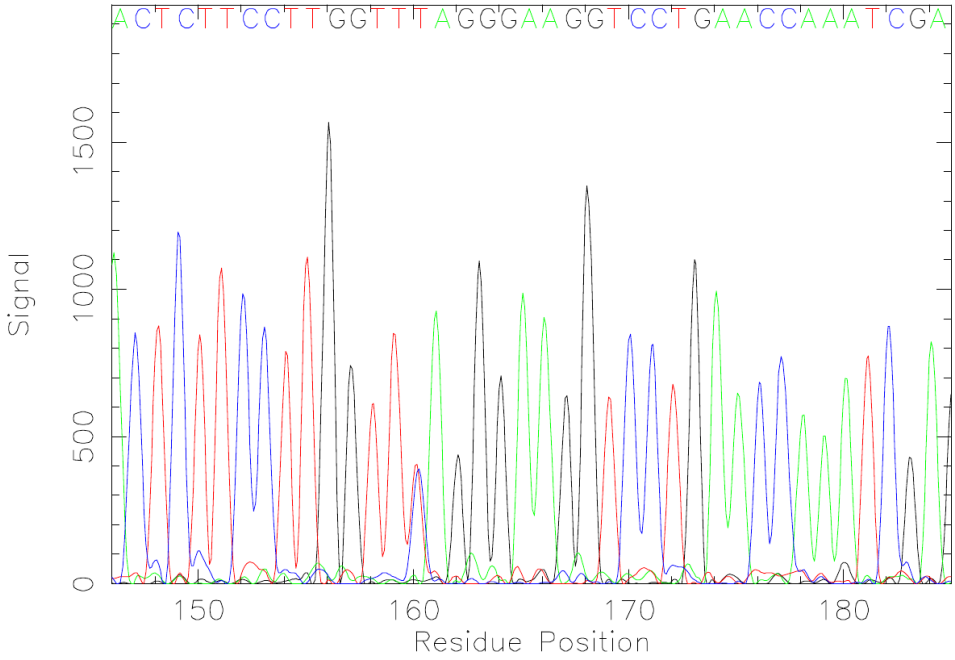


Figure 2.2. Chromatogram for hsa-miR-659-5p candidate SNV

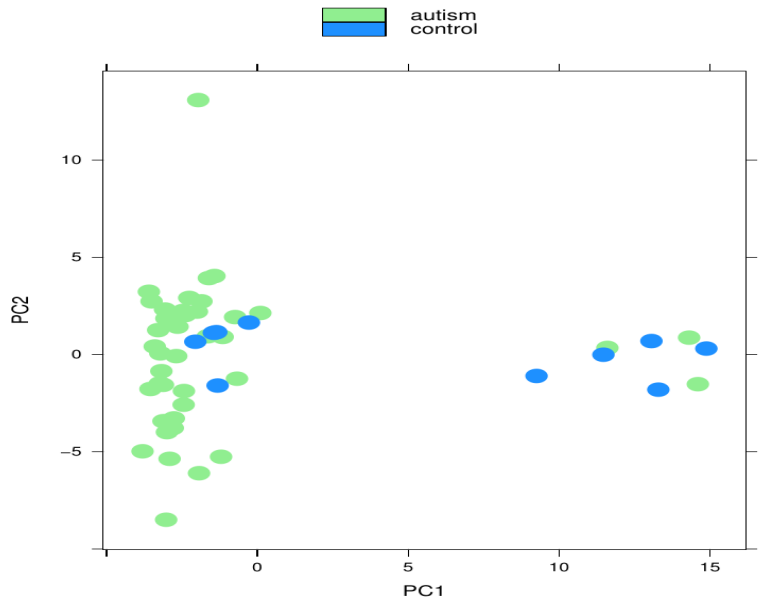
Chromatogram displayed by Abiview for Sanger sequencing results of validation of SNV at chromosome 22, position 38243748. The position of the candidate SNV on the sequenced read was residue 160.

The Sanger sequencing results suggest that the called rare seed region variants were not present within our sample. This poor validation could possibly reflect accumulation of variants within the lymphoblastoid cell lines, which were not present in the original genomic DNA (Yuen et al., 2016). However, assessment of the read depths supporting some of the variants also suggests that sequencing error could be responsible, with some such as the candidate variant on hsa-mir-4301 having only 3 supporting reads. This suggests that the Phred score threshold of 17.0 that we used was too low to produce reliable SNV calls. An adjustment from 17.0 to 30.0 would change the prior probability of error from 0.02 to 0.001. Filtering to require Phred scores above 30.0 yielded no seed region variants, which suggests that poor quality and low numbers of sequencing calls was the most likely cause of the false positives.

2.4.4. MicroRNA expression profiles

HTSeq-count was used to count the reads aligned to annotated mature microRNAs. These counts were then used in DESeq2 to assess differential expression (Anders & Huber, 2010). Raw counts for each sample and mature microRNA combination were modelled and normalised using a negative binomial distribution in DESeq2. We then used DESeq2 to normalise counts by regularised log transformation due to the large difference in read counts between samples. This normalised count data was then used for principal components analysis. We visualised the clustering of samples based on expression profiles using the first two principal components for microRNA expression levels (Figure 2.3).

(a)



(b)

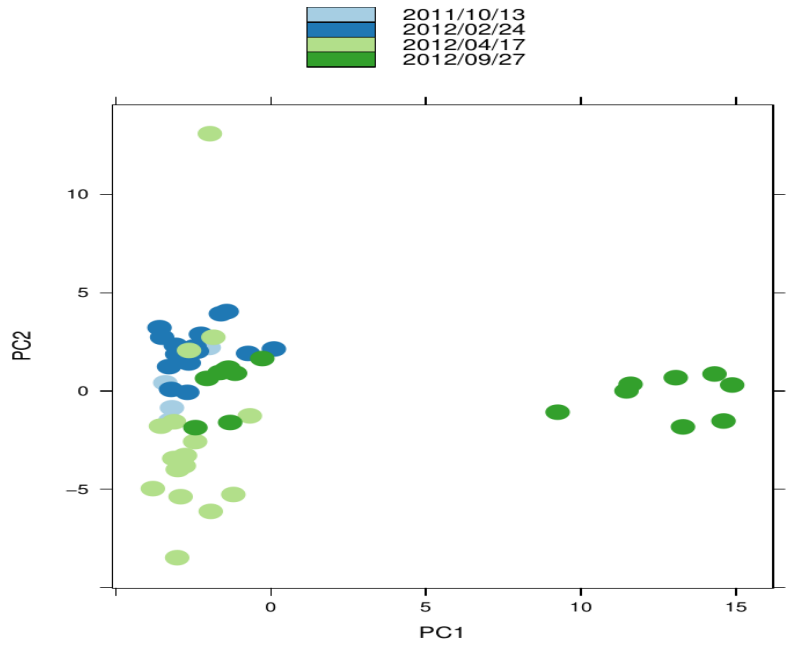


Figure 2.3. PCA plots for sample expression profiles

Samples plotted based on the first two principal components of normalised microRNA expression profiles. Spots indicating sample positions in the space of the first two principal components are (a) coloured green for autism samples and blue for controls and (b) coloured by sequencing batch.

The principal components analysis indicates a clear clustering of samples into two groups, with one containing half of the control samples, and the other having the majority of autism samples (39 out of 42).

We investigated for possible effects from the sequencing batch in which RNA samples were processed. Samples were sequenced in four groups, labelled for their date, on 2011/10/13, 2012/02/24, 2012/04/17 and 2012/09/27. We performed principal components analysis as described above, with samples labelled by batch (Figure 2.3). Labelling samples by sequencing batch shows that all of the members of the second cluster were sequenced in a single batch on 2012/09/27. In order to investigate clustering of samples more formally without visual inspection, we performed complete-linkage hierarchical clustering based on the normalised expression profiles and calculated pairwise similarity between samples (Figure 2.4).

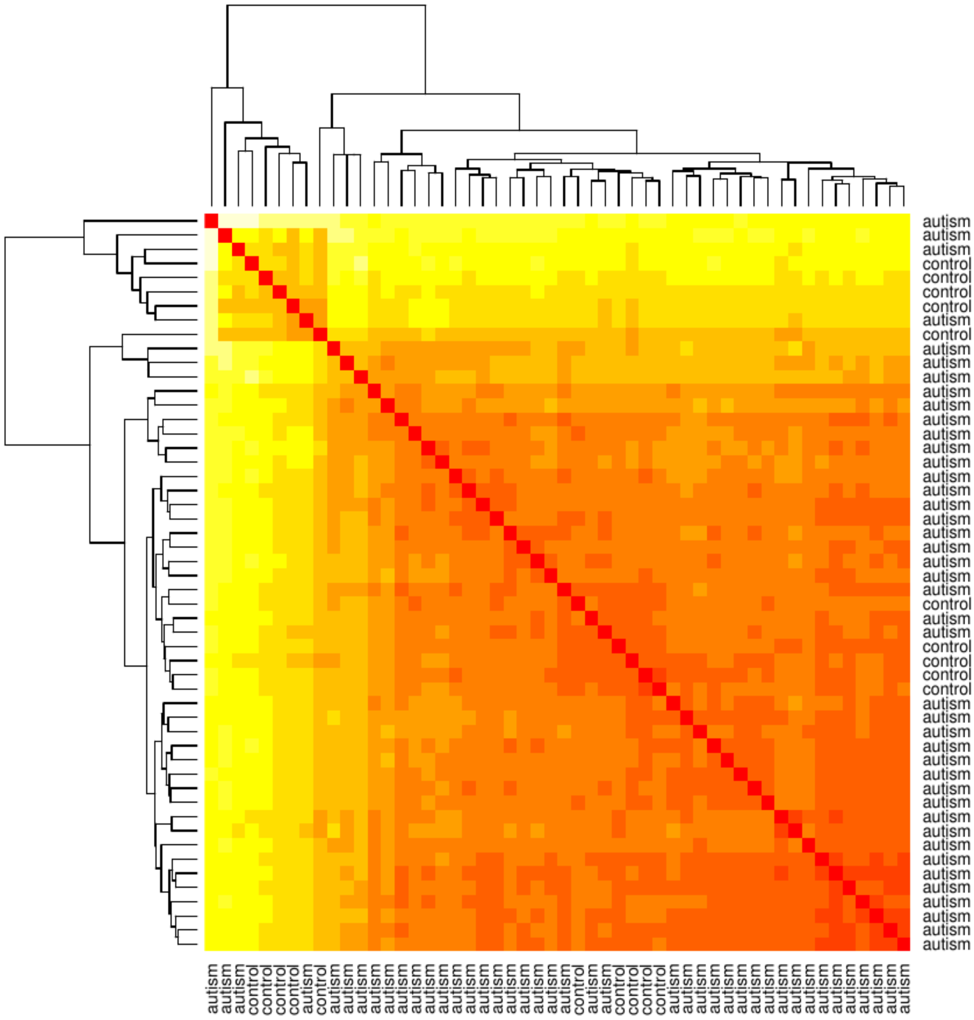


Figure 2.4. Sample clustering and heatmap

Complete linkage hierarchical clustering was performed based on pairwise similarity in expression profiles between samples. Profiles were normalised by regularised log transformation, and distance between samples was calculated by Euclidean distance across the array of microRNA normalised expression counts. The dendogram produced by clustering is displayed, and heatmap ordered by this clustering. Colour represents relative similarity of sample pairs, with white, then yellow the least similar and red indicating identical expression (self-self comparisons).

Complete-linkage clustering shows a moderate grouping of controls together, in two different tightly grouped sets. Two squares of adjacent, similar samples on the heatmap (Figure 2.4), at the very left and more weakly in the centre, contain 5 control samples each. While clustering based on complete microRNA expression profile does not clearly separate all of the control samples, similar results have been reported previously (Hicks & Middleton, 2016). Therefore, we investigated whether individual microRNAs may have consistent dysregulation between autism and control samples. Such variations in autism may indicate a specific alteration in pathway regulation.

2.4.5. Differential expression

We performed differential expression analysis with DESeq2 for the read counts supporting mature microRNAs in autism and control samples (Anders & Huber, 2010). When normalising read counts, we handled samples with outlier expression for a microRNA as measured by a Cook's distance indicating a single sample affected resultant statistics excessively. Those samples were included in the differential expression analysis, but the normalised read count for that microRNA in the sample was set to the mean of that microRNA across all other samples. We followed this approach as a conservative way to handle outliers, without increasing the chance of calling significant differential expression. However, we also repeated the differential expression analysis without trimming outlier read counts to the mean, which produced an identical list of differentially expressed microRNAs. In total, we identified 24 mature microRNAs with differential expression between autism and control in our experiment (Table 2.4).

Table 2.4. Differentially expressed microRNAs

MicroRNA	Fold Change in Autism (log 2)	P-value	Adjusted P-value
hsa-let-7c-5p	-2.115	0.0000	0.0002
hsa-miR-181b-3p	1.600	0.0000	0.0002
hsa-miR-532-5p	1.551	0.0000	0.0009
hsa-miR-3609	-2.610	0.0000	0.0010
hsa-miR-155-3p	0.900	0.0000	0.0028
hsa-miR-642b-5p	1.313	0.0001	0.0033
hsa-miR-99a-5p	-1.890	0.0002	0.0084
hsa-miR-132-5p	1.393	0.0004	0.0124
hsa-miR-223-3p	1.189	0.0004	0.0124
hsa-miR-32-5p	1.114	0.0004	0.0124
hsa-miR-3613-5p	2.043	0.0004	0.0124
hsa-miR-423-5p	-1.302	0.0003	0.0124
hsa-miR-188-5p	0.969	0.0006	0.0158
hsa-miR-223-5p	1.414	0.0008	0.0186
hsa-miR-6723-5p	1.995	0.0008	0.0186
hsa-miR-222-5p	1.176	0.0011	0.0241
hsa-miR-186-5p	-0.764	0.0014	0.0266
hsa-miR-34c-5p	-2.347	0.0014	0.0266
hsa-miR-642a-5p	1.124	0.0015	0.0266
hsa-let-7b-5p	-1.573	0.0017	0.0291
hsa-miR-138-1-3p	1.947	0.0020	0.0333
hsa-miR-125b-5p	-1.640	0.0025	0.0370
hsa-miR-30d-3p	1.388	0.0024	0.0370
hsa-miR-214-3p	-1.795	0.0034	0.0488

List of 24 microRNAs identified as differentially expressed in autism. Log2 fold change gives the change in autism versus control expression levels. FDR adjustment is done by Benjamini-Hochberg procedure, selecting those below 0.05 gives an FDR of 5%.

Of the 24 microRNAs identified with differential expression at FDR of 5%, 15 showed an increase in expression in autism, from approximately doubled (hsa-miR-155-3p) to quadrupled (hsa-miR-3613-5p). Of those microRNAs with reduced expression in autism, hsa-miR-3609 had the greatest change, at 16% of control expression in autism. Of the microRNAs which we found differentially expressed, 5 have previously been reported as differentially expressed in autism. Upregulation of hsa-miR-181b was previously reported in lymphoblastoid cell lines (Ghahramani Seno et al., 2011). As well as previous reports in several other neurological disorders (Edbauer et al., 2010; Hara et al., 2017), hsa-miR-132-5p was reported as differentially expressed in autism in three other studies, in brain, saliva and LCL (Abu-Elneel et al., 2008; Sarachana et al., 2010; Talebizadeh et al., 2008). Interestingly, these reports had conflicting directions of dysregulation, with one reporting down-regulation (Abu-Elneel et al., 2008) and two up-regulation matching our results (Sarachana et al., 2010; Talebizadeh et al., 2008). More recently, upregulation of hsa-miR-223-3p has been reported (Y. E. Wu et al., 2016). Downregulation of hsa-miR-32-5p was reported previously, in contrast to significant upregulation in our results (Hicks et al., 2016). Finally, matching downregulation of hsa-miR-186-5p was reported in saliva previously (Sarachana et al., 2010). In general, previous tests of microRNA expression in various autism samples have largely different lists of differentially expressed microRNAs. We did not find differential expression for the three microRNAs (hsa-miR-23a, hsa-miR-146a and hsa-miR-106b) which were found dysregulated previously in at least three other studies (Hicks & Middleton, 2016).

Visual inspection of normalised count data for the 24 differentially expressed microRNAs in some cases showed clear bimodal distributions (Figure 2.5). Others showed more consistent expression across samples.

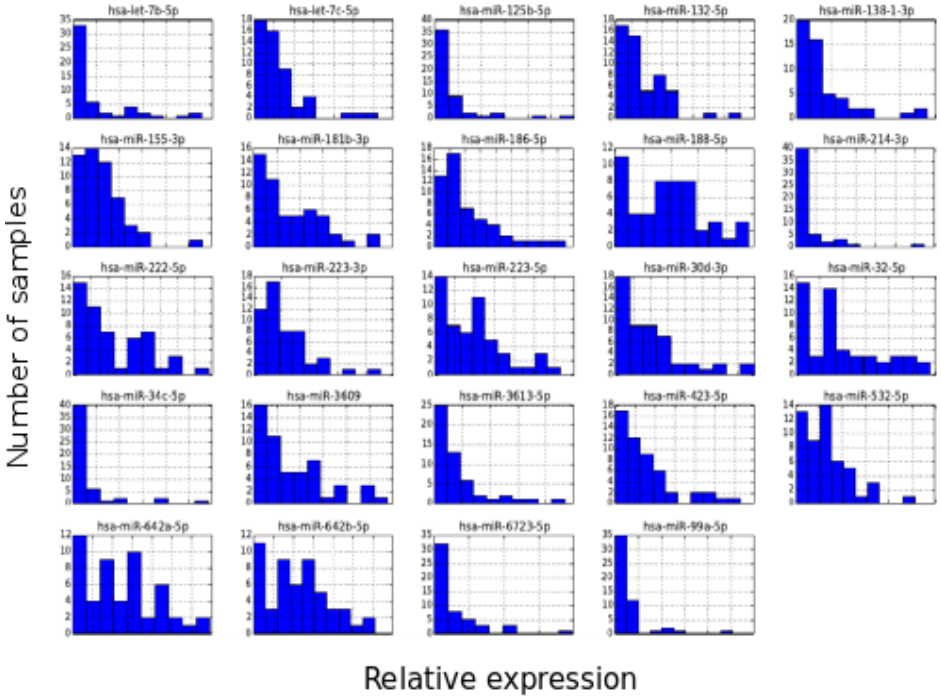


Figure 2.5. Relative normalised expression for differentially expressed microRNAs

Histograms of normalised expression for 24 differentially expressed microRNAs. Read count data was normalised by a simple division with total sample microRNA aligned reads. The Y-axis counts the number of samples with expression within each decile. X-axis labels are not shown for simplicity as sample normalised counts vary from max 0.112 (hsa-miR-155-3p) to max 0.000003 (hsa-miR-34c-5p).

2.4.6. Detection and analysis of isomiRs

We developed a script to count expression levels by starting nucleotide position as an alternative to HTSeq-count which would allow us to distinguish between isomiRs (Anders & Huber, 2010). From the aligned SAM input, reads within 3 bases of a mature microRNA annotation and on the same strand were grouped by 5' start position. Alternative 3' isomiRs were ignored due to the unreliable end point left by sequential trimming in SeqTrimMap and known loss of quality at the ends of SOLiD reads (Marco & Griffiths-Jones, 2012). This procedure identified all distinguishable different mature microRNA products in the sequencing data - some of which were genuine isomiRs of the same microRNA, and some of which were simply mature products in different microRNAs or alternative arms of the same microRNA. For simplicity, and to indicate that these annotations take 5' start position into account in contrast to standard miRBase annotations, we will use the general term 5' isomiR for such products. We only considered 5' isomiRs which passed the threshold of at least 3 reads in at least 3 samples, of which we observed a total of 2,494. Counts generated by this script were input into DESeq2 and standard analysis as above yielded 24 differentially expressed 5' distinguishable transcripts between autism and control (Table 2.5).

Table 2.5. Differentially expressed 5' isomiRs

5' isomiR Parent miRNA	Fold Change (log2)	P-value	Adjusted P-value
hsa-mir-155	3.923	0.0000	0.0000
hsa-mir-181b-1	1.642	0.0000	0.0003
hsa-let-7c	-2.030	0.0000	0.0008
hsa-mir-532	1.582	0.0000	0.0012
hsa-mir-132	1.699	0.0000	0.0035
hsa-mir-150	2.143	0.0000	0.0038
hsa-mir-532	1.405	0.0001	0.0060
hsa-mir-4425	2.535	0.0001	0.0060
hsa-let-7a-1	-1.860	0.0001	0.0060
hsa-mir-142	2.332	0.0001	0.0067
hsa-mir-223	1.803	0.0001	0.0090
hsa-mir-188	1.016	0.0002	0.0123
hsa-mir-142	2.109	0.0002	0.0128
hsa-mir-3613	2.078	0.0003	0.0136
hsa-mir-99a	-1.798	0.0004	0.0150
hsa-mir-32	1.144	0.0003	0.0150
hsa-mir-191	1.726	0.0005	0.0186
hsa-mir-23a	1.963	0.0005	0.0201
hsa-mir-223	1.479	0.0006	0.0205
hsa-mir-423	-1.253	0.0006	0.0205
hsa-mir-342	-1.395	0.0006	0.0205
hsa-mir-30d	2.293	0.0012	0.0355
hsa-mir-222	1.185	0.0012	0.0358
hsa-mir-660	0.607	0.0016	0.0452

List of 5' isomiRs identified as differentially expressed in autism at FDR 0.05. 5' isomiRs are defined by 5' start position and belong to an annotated pre-microRNA. Log2 fold change in autism, p-value for differential expression and Benjamini-Hochberg adjustment are also shown.

This analysis of differential expression produced a similar set of differentially expressed transcripts as that based on overlap with known mature microRNAs. Apart from minor changes in ordering and p-value, there were 7 omissions from the set of differentially expressed 5' isomiRs: miR-3609, miR-642b-5p, miR-6723-5p, miR-186-5p, miR-34c-5p, miR-642a-5p and miR-125b-5p. The absence of 5' isomiRs near to these annotated microRNAs with reliable differential expression suggests that reads supporting those original differential expression calls were not consistently aligned with a single start site. Conversely, we observed interesting signals of differential expression for 5' isomiRs located within the pre-microRNAs mir-150, mir-521, mir-4425, let-7a-1, mir-142, mir-191, mir-23a, mir-342 and mir-660 which were not previously identified using standard read counting. Of these, mir-142 is particularly interesting, as it was highlighted in a previous study showing differential expression and hypomethylation in autism brain samples (Mor et al., 2015). The microRNAs miR-191-5p and miR-342 were also previously identified, although strangely miR-191-5p was reported with differential expression in both directions previously (Hicks et al., 2016; Sarachana et al., 2010). Most valuable is the recovery of a signal for differential expression in mir-23a, which is reflected in repeated observations of differential expression of mature products of this microRNA in previous studies (Hicks & Middleton, 2016). Failure to identify this microRNA in the standard differential expression analysis may have been due to noise from poorly aligned overlapping reads not truly corresponding to the mature microRNA.

We investigated changes in the processing of the dominant versus other isomiRs for each mature annotated microRNA in autism and control samples. In order to do this, we performed pairwise comparisons at each annotated mature microRNA. We counted the number of reads supporting the dominant 5' isomiR and the number of reads supporting other 5' isomiRs in each sample. We then calculated a ratio between the counts for each sample and for each annotated microRNA. Samples without reads for either isomiR were discarded in the tests. We then performed a two-sided Mann-Whitney U-test on these ratios between autism and control samples for each

microRNA.

Before multiple testing correction was applied, we obtained 52 mature microRNAs with significantly different isomiR usage in autism and controls (Supplementary Table 2.1). However, none of these were significant with false discovery rate 5% after applying Benjamini-Hochberg adjustment. Therefore, the relative expression of different 5' isomiRs does not appear to change significantly between autism and control.

2.4.7. Arm usage changes in autism

We examined the usage of the 5p and 3p arms for each microRNA, to investigate whether arm usage differed for any microRNAs between autism and control samples. We aggregated all of the reads supporting 5' isomiRs on the correct strand within 5 bases of the start sites of annotated mature microRNAs. As with the 5' isomiR dominant usage analysis, for each primary microRNA, we calculated the ratio of the number of 5p reads against the number of 3p reads. Samples without reads for either arm were discarded in the analysis for each microRNA. We performed a two-sided Mann-Whitney U-test on ratios for each microRNA to test for significant difference in usage between autism and control samples.

Without multiple testing correction, we found 38 primary microRNAs with significantly different arm usage in autism. After Benjamini-Hochberg adjustment, we identified 8 pre-microRNAs with significantly different arm usage (Table 2.6).

Table 2.6. Significant arm usage changes in autism

MicroRNA	Mann-Whitney U-test P-value	Adjusted P-value
hsa-mir-3613	0.000	0.025
hsa-mir-505	0.001	0.048
hsa-mir-200b	0.001	0.048
hsa-mir-642a	0.001	0.048
hsa-mir-92a	0.001	0.048
hsa-let-7f	0.002	0.048
hsa-mir-181b	0.001	0.048
hsa-mir-146a	0.001	0.048

List of microRNAs with significantly different usage of 5p and 3p arms between autism and control samples. Reads were counted as supporting an arm where they aligned to start within 5 bases of the miRBase annotated mature sequence on the correct strand.

We previously found hsa-mir-181b-3p involved in differential expression between autism and control samples. The changing arm usage observed here may reflect that change, combined with consist expression of the 5p arm in the two groups. Most interestingly, hsa-miR-146a-5p has been identified as differentially expressed in other studies, with all finding upregulation, in brain, LCL and olfactory stem cells (Mor et al., 2015; Nguyen et al., 2016; Talebizadeh et al., 2008). Our standard differential expression analysis did not find any significant change in our samples, however examination of arm usage supports changes in regulation in autism.

2.4.8. Enrichment analysis

Taking the set of 24 differentially expressed microRNAs in autism from our results, we explored the set of genes predicted to be targeted. We used multiple target prediction algorithms to generate alternate predicted gene sets. We used a database of targets downloaded for humans from DIANA microT-CDS (Reczko et al., 2012). We ran miRanda version 3.3a requiring free energy under 20kcal per mol and target score above 155 with the target database Ensembl release 75 (Enright et al., 2003). We downloaded target data from the PITA Targets Catalog v6 (Kertesz et al., 2007). Finally, we also ran the seedVicious algorithm to directly predict gene targets based on seed matching, again using human genes from Ensembl release 75 (Marco, Hui, Ronshaugen, & Griffiths-Jones, 2010). We counted a total of 11,197 targets for our differentially expressed microRNAs predicted by miRanda. We used an intersection of the remaining three prediction tools (PITA, DIANA microT, seedVicious), as an alternative approach, although noting that this may not necessarily improve accuracy (Ritchie et al., 2009). We obtained 8,272 targets for the 24 differentially expressed microRNAs (Ritchie et al., 2009). Such large target sets and concerns over the ability of standard functional enrichment analysis to produce valid results motivated development of alternative enrichment analysis tools, as discussed in the following chapter.

We used the functional enrichment analysis tool miEAA to investigate the properties of the targets of these microRNAs (Backes et al., 2016). We input the list of differentially expressed microRNAs for over-representation analysis, examining categories for pathways (using miRWalk target predictions), diseases, gene ontology and Tarbase experimentally validated convergent targeting (Papadopoulos et al., 2009). Only a single category was enriched after multiple testing correction in these collections - experimentally confirmed targeting of gene KIAA0141 by three of the differentially expressed microRNAs. This gene, a death ligand signal enhancer, does not have an obvious connection to autism susceptibility.

We developed an algorithm to examine enrichment of predicted target sets for differentially expressed microRNAs, as described later in this thesis (Bleazard et al., 2015). We tested our 24 differentially expressed microRNAs for enrichment of gene ontology classes using basic miRanda target predictions, using a multihit miRanda target model, and using target predictions based on the intersection model. No significant categories were returned after multiple testing correction at FDR 5%, except for two in the more sensitive multihit model (FDR 1%). These were GO:0032023 trypsinogen activation and GO:0060212 negative regulation of nuclear-transcribed mRNA poly(A) tail shortening. Although gastrointestinal disorders are often present in autism, these gene ontology groups do not have obvious connections to gene networks thought to be involved in autism susceptibility (Lai et al., 2013). We used our algorithm to perform a more directed assessment of functional enrichment specifically for known autism-related genes as annotated by AutDB (Basu et al., 2009). Using a set of 537 genes annotated with some evidence for connection to autism, we ran our enrichment analysis multihit algorithm. The collection of autism genes was not significantly over-targeted by the differentially expressed microRNAs ($p=0.19$).

2.4.9. Machine learning for prediction of autism status

Following the hypothesis that microRNA profiles may provide an effective classifier for autism status, we attempted to use machine learning to build a prediction model. We pre-processed count data obtained from HTSeq-count, normalising read counts with simple division with the total reads aligned in each sample (Anders & Huber, 2010). Our dataset contained 52 records, of which 10 were labelled control, and 2,576 features, which were mature microRNAs with non-zero aligned reads. We explored correlations between normalised read counts for a restricted feature set of microRNAs highlighted for differential expression, arm switching and isomiR expression above (Figure 2.6).

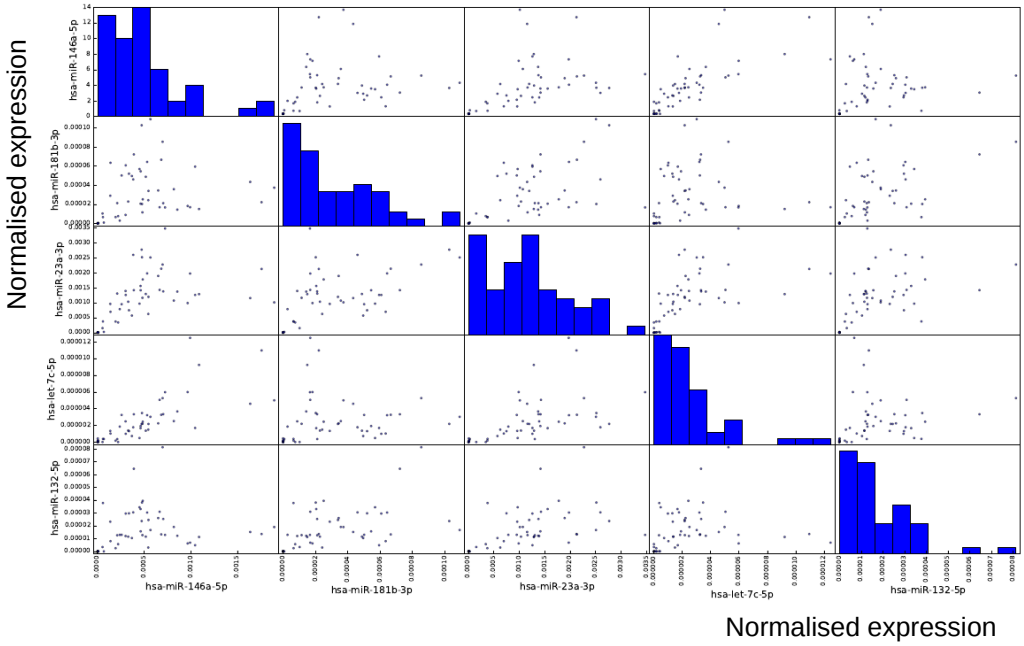


Figure 2.6. Normalised sample expression for pairs of dysregulated microRNAs

We plot histograms on the diagonal boxes of normalised read counts for 5 microRNAs of interest described above: hsa-miR-146a-5p, hsa-miR-181b-3p, hsa-miR-23a-3p, hsa-let-7c-5p and hsa-miR-132-5p. For boxes off the diagonal, scatter plots show comparisons of sample expression levels for each pair of microRNAs.

Scatter plots show strong correlations in normalised read counts for microRNAs of interest. The two microRNAs selected here which were not identified as differentially expressed (hsa-miR-146a-5p and hsa-miR-23a-3p) interestingly show strong pairwise correlation in expression and also strong correlation with other differentially expressed microRNAs.

Using the normalised read counts, we formatted data for input to the machine learning module scikit-learn (Hall et al., 2009). Due to the very small number of control samples available, we relied on 10-fold cross validation to produce assessments of model accuracy while still providing sufficient control examples for training. We used a logistic regression model with a regularizing factor of $1.0x10^{-5}$ and l1 norm to avoid over-fitting. We also performed this regression with a scikit-learn default factor of 1, but obtained a model without control calls. Coefficients of the 2,576 microRNA expression levels were optimised for predictive accuracy. Random 10-fold cross validation of this model produced a confusion matrix with total accuracy of 81% (Table 2.7).

Table 2.7. Machine learning confusion matrices and performance

Logistic regression	**Autism prediction**	**Control prediction**
Autism test sample	35	7
Control test sample	3	7
Accuracy	81%	
K-nearest neighbours	**Autism prediction**	**Control prediction**
Autism test sample	37	5
Control test sample	8	2
Accuracy	75%	
Bayes classifier	**Autism prediction**	**Control prediction**
Autism test sample	34	8
Control test sample	4	6
Accuracy	77%	
Support vector machine	**Autism prediction**	**Control prediction**
Autism test sample	38	4
Control test sample	6	4
Accuracy	81%	

Predictions of 10 models each generated by leaving out an alternate tenth of the sample data were assessed against the remaining data, and outcomes aggregated. Results are shown for logistic regression, 5-nearest neighbours, naive Bayes classifier and support vector machine models.

The asymmetry in sample counts for autism and control means that general accuracy estimates are unlikely to reflect performance on more balanced samples. For this reason, the sensitivity and specificity for controls are likely the most informative measures of performance of models. Here, we observe that autism calls had 83% sensitivity and 92% specificity. However, sensitivity for controls was only 70% and specificity for control calls was 50%. We also attempted logistic regression after a preprocessing step on the input data to transform each microRNA's normalised counts to have zero mean and unit variance. However, this reduced performance on test data, with 40 autism samples predicted autism status, yielding control specificity 18%.

We explored several alternative models for machine learning due to limited performance of logistic regression. We implemented a k-nearest neighbours vote classifier, predicted autism status for test sets in 10-fold cross-validation based on a sample's 5 nearest samples in the training set by Euclidean distance between microRNA expression profiles. This model produced an accuracy of 75% (Table 2.7).

Sensitivity for autism was 88% using this model, however sensitivity for control was only 20%. Specificity for autism was 82% and specificity for control calls was 71%. We explored using decision trees as models for classification, but did not obtain better than 50% sensitivity for control samples, likely due to the large number of features with few samples. Similarly, using a random forest model was impractical due to the small number of control samples, leading to branch decisions without any controls in their training set. We implemented a naive Bayes classifier, assuming that the likelihood function for each microRNA expression was independent and Gaussian, although noting that these assumptions were not necessarily valid. Testing with 10-fold cross-validation, we obtained an accuracy of 77% (Table 2.7).

The Gaussian naive Bayes classifier produced sensitivity of 81% for autism status, 89% specificity for autism calls, 60% sensitivity for control samples

and 57% specificity for control calls. We constructed a support vector machine classification model using a linear kernel. Using data not scaled for unit variance, the model did not learn to predict control for any samples in testing. Therefore we used scaled input data with mean zero and unit variance. Testing this model using 10-fold cross-validation yielded an accuracy of 81% (Table 2.7).

The support vector machine classifier had specificity for autism calls of 86%, sensitivity for autism of 90%, specificity for control calls of 50% and sensitivity for control samples of 40%. Due to selection of alternative parameters for various models, it is likely that estimates of accuracy in our 10-fold cross-validation tests are higher than would be obtained on new data. We tested whether results were robust when a smaller training set was provided, with 38% of the samples reserved for validation. Building a model with 62% of the samples using logistic regression, we obtained a classifier with bias for calling control samples, with sensitivity for controls of 100%, but sensitivity for autism of 59% and only 30% specificity for control calls.

The best model developed here, logistic regression, achieved 81% accuracy, but specificity for control calls of only 50%. In general, noting the high accuracy generally demanded of classifiers for conditions such as autism, such performance is not sufficient.

2.4.10. Detection of A-I editing

In order to investigate features of the transcriptome beyond expression levels, we explored post-transcriptional alterations to microRNA sequences. Previous studies have reported A-I editing in human microRNAs, including 19 specific locations where it had been previously observed in human brain tissue (Alon et al., 2012). We examined raw aligned sequence data for replication of an editing signature in our own results. We filtered reads to only include those with a Phred score of at least 30, and then counted the number of G-supporting reads at those locations against the count of A-

supporting reads. This threshold was used as it gave a probability of error of p=0.001 for each sequenced read at that position. We tested against the null hypothesis of sequencing error producing counts of G-supporting reads under a binomial distribution. We found two positions with highly significant A-I editing out of 19 prior positions tested ($p < 4x10^{-5}$). These were in miR-99a-5p on chromosome 21 at position 17911421 and in miR-589-3p on chromosome 7 at position 5535483. Interestingly, hsa-miR-99a-5p was among the microRNAs which we found differentially expressed in autism. Although we could not find differential A-I editing at this site in autism, the consistent post-transcriptional modification at this loci may indicate complex unknown regulation. We also ran our algorithm across all adenine bases in microRNAs throughout the genome, but did not identify significant A-I editing after correction for multiple testing.

2.4.11. Validation of differential expression

We performed qRT-PCR for validation of the set of differentially expressed microRNAs called from our sequencing data. We selected the top seven differentially expressed microRNAs, which were also the subset that had a false discovery rate of 1% according to Benjamini-Hochberg adjustment. To this test set, we also added miR-223-3p and miR-223-5p and hsa-let-7a-3p, as these also showed different expression patterns in autism and were of interest due to the literature or explored transcriptome features in this study. We used Taqman Advanced microRNA Assays designed to target these microRNAs. A custom Taqman Advanced microRNA Assay was designed by Applied Biosystems for hsa-miR-181b-3p (Table 2.8). We were unable to test two of the seven differentially expressed microRNAs at FDR 1% due to the assays not arriving in time for this study.

Table 2.8. Taqman Advanced microRNA assays

MicroRNA	Mature microRNA Reference	Status
hsa-miR-99a-5p	AACCCGUAGAUCCGAUCUUGUG	Test
hsa-miR-181b-3p	Custom target assay	Test
hsa-miR-532-5p	CAUGCCUUGAGUGUAGGACCGU	Test
hsa-let-7c-5p	UGAGGUAGUAGGUUGUAUGGUU	Test
hsa-miR-3609	CAAAGUGAUGAGUAAUACUGGCUG	Test
hsa-miR-223-3p	UGUCAGUUUGUCAAAUACCCCA	Test
hsa-miR-223-5p	CGUGUAUUUGACAAGCUGAGUU	Test
hsa-let-7a-3p	CUAUACAAUCUACUGUCUUUC	Test
hsa-miR-186-5p	CAAAGAAUUCUCCUUUUGGGCU	Endogenous control
hsa-miR-191-5p	CAACGGAAUCCCAAAAGCAGCUG	Endogenous control
hsa-miR-320a	AAAAGCUGGGUUGAGAGGGCGA	Endogenous control

List of microRNAs for which qRT-PCR was used to validate differential expression. Three microRNAs were selected as endogenous controls for normalisation.

We selected three microRNAs as endogenous controls, with hsa-miR-186-5p and hsa-miR-191-5p being standard for use as controls and hsa-miR-320a showing particularly stable expression within our small RNA sequencing data. We confirmed stable expression for all three controls within the sequencing data with a script testing for range in normalised read counts and no outlier samples.

We performed optimisation tests for each probe and obtained successful amplification (Ct below 31, technical replicate standard deviation below 0.3). We divided samples into 12 plates with three replicates for each sample-probe pair, negative controls for each microRNA probe and three endogenous controls within each plate for normalisation. In total we performed 3,726 qRT-PCR reactions. Ct values were calculated by Quantstudio 12K Flex automatic algorithm for each test. For each microRNA probe-sample pair, we filtered those with outlier values among the three technical replicates, or which failed quality filters (Methods). We repeated the test process for reactions failing quality tests, and then excluded from further analysis those which failed successful amplification. This left a total of 208 delta Ct scores for testing for significant differences in groups. We used an average of the three replicates for each probe-sample pair, and normalised by subtracting the average Ct value for the three endogenous controls for the sample sample. This delta Ct value for each sample for each probe was then used in a T-test against the null hypothesis that delta Ct values were the same for autism and control samples. Of the tested microRNAs, we were able to obtain successful validation of 7 (Table 2.9).

Table 2.9. qRT-PCR validation results

MicroRNA	P-value
hsa-miR-99a-5p	0.0010
hsa-miR-181b-3p	0.0126
hsa-miR-532-5p	0.0010
hsa-miR-577	0.0069
hsa-miR-3609	0.0235
hsa-miR-223-3p	0.0330
hsa-miR-223-5p	0.0444
hsa-let-7a-3p	0.1245

T-test p-values reported for microRNAs for which amplification and filtering were successful. Seven of the tested microRNAs passed validation steps.

The direction of differential expression in the qRT-PCR tests matched with the direction of change in the RNA-seq dataset. We generated box-plots for delta Ct values in autism and control for each microRNA (Figure 2.7).

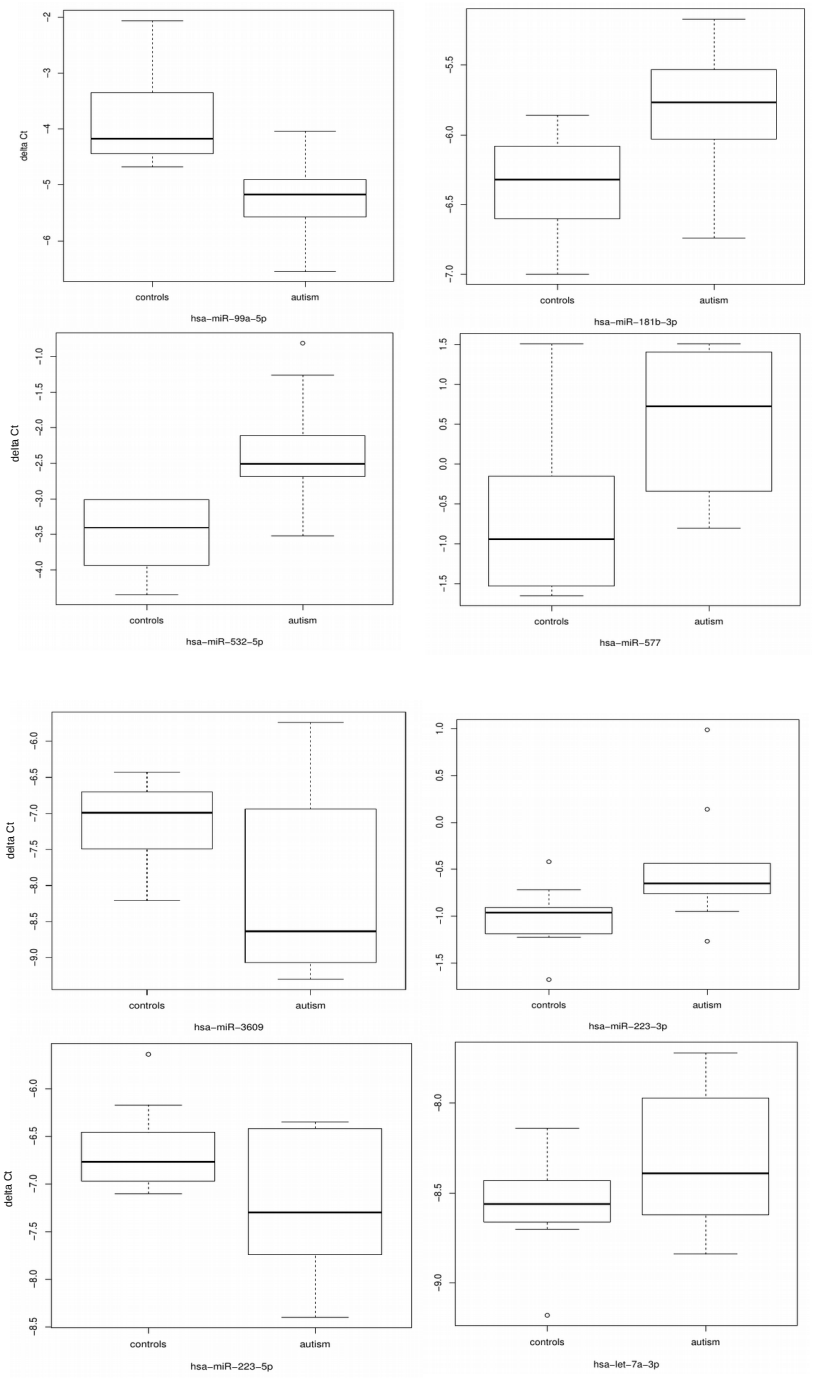


Figure 2.7. Delta Ct values for validated microRNAs

Boxplots for delta Ct values for control and autism samples for each microRNA probe. Delta Ct values are given on the y-axis.

The qRT-PCR validation demonstrates that the differential expression called by small RNA sequencing data analysis is largely robust. However, several of the highly differentially expressed microRNAs could not yet be confirmed, and several outlier samples were present affecting the test statistic. Further testing may be necessary for more reliable confirmation of differential expression.

2.4.12. Subclassification by phenotype

We investigated how more detailed assessment of individual phenotype could provide informative classification of samples and improve understanding of the role of microRNAs in different core impairments in autism. We first annotated the sequenced samples in our study based on IMGSAC phenotype classification criteria (IMGSAC, 2001). Our sample contained 23 individuals with case type 1, defined as those with International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) autism criteria and clinical diagnosis, meeting ADI-R and ADOS autism criteria, IQ above 35 and history of language delay. Our sample contained 13 individuals with case type 2, defined as those passing ICD-10 criteria for autism, Asperger's syndrome, atypical autism or pervasive developmental disorder (PDD), passing ADI-R autism criteria for 2 domains and borderline for a third, meeting criteria for at least PDD on the ADOS, IQ above 35 and some abnormal development before the age of 3. One individual in our study was of case type 3, with evaluation matching those in case type 2, but without ADOS assessment for PDD. Three individuals in our sample were of case type 4, meeting ADI-R and ICD-10 criteria for autism, and matching criteria for PDD on the ADOS, and including those with IQ below 35. A remaining two samples could not be assigned to these classes unambiguously, but had diagnosis for autism and pervasive developmental disorder respectively.

We investigated the clustering of our samples within the larger IMGSAC dataset based on autistic individual phenotypes. We used the full collection

of IMGSAC samples with complete ADI-R feature scores for this analysis, to allow a better understanding of the relative phenotypes of sequenced individuals in the context of a wider collection of those diagnosed. This included 1,351 individuals with scores for 205 attributes. We scaled these responses for machine learning, setting each feature mean to zero and transforming to give unit variance. We performed principal components analysis and transformed our data to a reduced dataset with only the first two principal components for each sample. Visual inspection of the reduced dataset showed clear clustering into three groups. We used k-means clustering to formally separate the dataset, requiring 3 clusters, and plotting decision boundaries in a PCA plot (Figure 2.8a).

We associated IMGSAC phenotype IDs with sample IDs used in our RNA-seq study, and were able to match full ADI-R feature data to 27 individuals. We further divided these samples into 10 annotated for PDD and 17 other autistic individuals (Figure 2.8b).

We found that the IMGSAC defined clusters were not supported by separation of pervasive developmental disorder samples. Similarly, we annotated our samples based on type classifications (C1, C2, C3, C4) as described above and plotted with labels for this categorisation (Figure 2.8c).

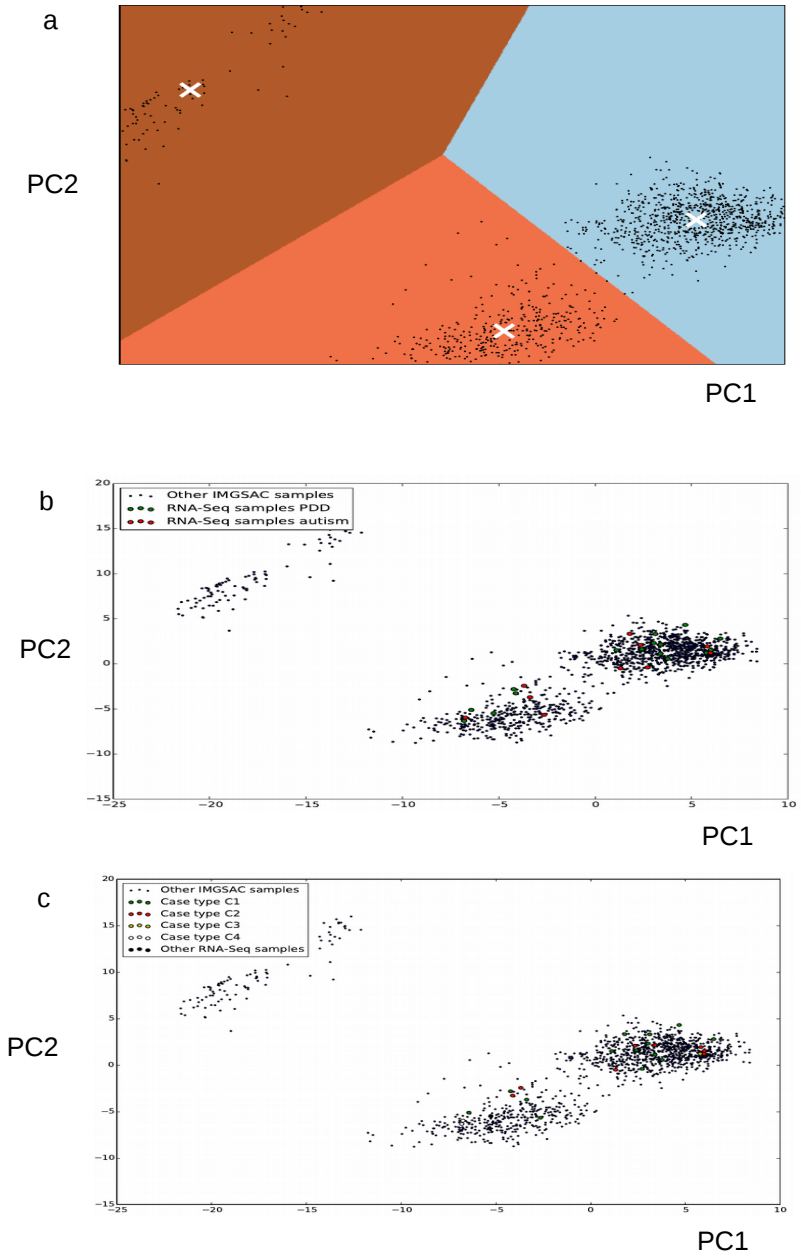


Figure 2.8. IMGSAC samples plotted by ADI-R PCA

Plot of IMGSAC samples (black spots) with full ADI-R scores for 205 features available on the transformed feature space of the first two principal components after PCA (x and y axis respectively). Decision boundaries for three clusters determined by k-means clustering are shown in (a), with centroids marked by white crosses. (b) Samples which were included in small RNA sequencing in our project are labelled green and red based on classification for pervasive developmental disorder. (c) Samples are plotted as spots labelled for case type.

Case type again fails to support a clear separation for two of the clusters indicated by PCA. Previous studies have explored clustering of samples similarly, and using fitness of merit analysis and correspondence analysis argue for a model for ADI-R phenotypes with two factors and three classes as observed here (Georgiades et al., 2013). The factors previously identified were associated with (a) social communication deficits and (b) fixated interests and repetitive behaviours. If we map the three classes described previously to the clusters identified above, we interpret top-left on the PCA plot above as previously labelled Class 3, highest social communication impairment and highest scores for repetitive behaviour. The remaining two clusters would map to previously described Class 1, moderately high social communication impairment and low repetitive behaviours and Class 2, with less social communication impairment and moderately high scores for repetitive behaviours (Georgiades et al., 2013). These classes do not correspond directly to the more commonly used separation of autism with Asperger's syndrome and pervasive developmental disorder or case types noting test passes (Figure 2.8).

2.5. Discussion

In this work, we have presented results from investigations into microRNAs in autism through small RNA sequencing. Our most significant findings include 24 differentially expressed microRNAs, of which 19 are novel in autism, and 5 replicate previous findings. We also present novel methods for analysis of post-transcriptional regulation of microRNAs, with findings from these methods identifying dysregulation in autism for microRNAs previously implicated.

2.5.1. Study limitations

There are several limitations of our study which must be considered when drawing conclusions on this work. Separation of samples by expression profiles did not perfectly cluster autism and control samples separately. We

found that Sanger sequencing did not support rare seed region variant calls made with small RNA sequencing data. We validated differential expression of key microRNAs with qRT-PCR but did not provide a full evaluation of other microRNAs. Finally, limits to training sets meant that machine learning approaches were unable to provide desired accuracy.

2.5.1.1. Sample clustering

We explored the separation of autism and control samples through principal components analysis and hierarchical clustering. Plotting samples with the first two principal components showed a clear separation into two groups. The smaller group contained 5 out of 10 control samples, as well as 3 out of 42 autism samples. The probability of 5 or more out of 10 control samples being placed in such a cluster at random was 0.0002 (Fisher's exact test). However, labelling of samples by sequencing batch showed that all of the members of this cluster were sequenced in the 2012/09/27 batch, which was approximately evenly split across the two clusters. Therefore, it is not possible to be certain that the separation of samples by this clustering is due to autism status rather than batch effects. We explored clustering of samples in more depth through hierarchical clustering, which generated a tree with more complex branching. Three groups of highly mutually correlated samples were most clearly identified (Figure 2.3). These consisted of two clusters each with 5 control and 3 autism samples, and a third of highly correlated autism samples. In previous studies using RNA-seq analysis of microRNAs in autism, total expression profiles did not separate cases and controls (Ander et al., 2015; Y. E. Wu et al., 2016), although clustering with only differentially expressed selected RNAs naturally clusters effectively. There are several potential difficulties particular to autism microRNA studies, including autism heterogeneity and inconsistent and incomplete diagnosis in the population (Y. S. Kim et al., 2011). These features, and the complex nature of autism as a genetic disorder, may explain the failure to separate samples based on general microRNA profile.

2.5.1.2. Detection of rare variants

Sequencing statistics show most reads aligned successfully to the genome, although with some variation in read depth between experiments (ranging from 648,181 to 35,148,092 raw reads). Allowing trimmed reads with 2 mismatches and 5 multiple alignments meant that a large number of reads mapped to multiple locations, and that variants and lengths at the 3' end of reads were unreliable. Manual checking of SNVs within pre-microRNAs with Integrative Genomics Viewer (IGV) suggests that problems with the 3' end did not severely impact on variant calling. However, when we checked for the presence of rare called SNVs in seed regions, we found that none were supported by Sanger sequencing in genomic DNA. The subset of SNV calls which we tested were selected as those in seed regions where the prior probability of variants being present is low, due to strong microRNA conservation (Bartel, 2009). Autism samples for our study were also filtered to require multiplex patterns within families. This meant that *de novo* variants within an individual would not be the main driver for autism status. Therefore, variants within our samples affecting autism susceptibility must be present in other family members and potentially at a population frequency sufficient for detection and registry by dbSNP (Sherry et al., 2001). Notably, the confirmed dbSNP variant call which we tested by Sanger sequencing was validated. An alternative source of *de novo* variants affecting autism within our probands follows the hypothesis that multiplex autism is often driven by mothers carrying *de novo* or rare variants (Zhao et al., 2007). However, our search was unable to find such variants within microRNAs in our sample. For this reason, we expand our search to a larger dataset in a following chapter.

2.5.1.3. Biomarker limitations

We performed a broad range of machine learning methods, in an attempt to produce a reliable predictor of autism affection status based on microRNA expression profiles. Similar efforts have provided successful predictive

biomarkers for other conditions (Leidinger et al., 2013), and consistency of microRNA profiles across blood and brain samples provides support for the potential of this approach in autism (Talebizadeh et al., 2008). This work is motivated by evidence that early interventions, such as parent-mediated social communication intervention, may have a lasting positive impact if diagnosis can be achieved quickly and reliably (Pickles et al., 2016). It is uncertain how much of these positive impacts reflect general benefits from improved childcare. However, incorrectly diagnosing and treating non-autistic children would clearly be harmful. We suggest that, due to the financial costs, the loss of trust in diagnosis and potentially harmful treatment, molecular tests for autism must have specificity well above 99%. This would be equivalent to approximately only half of treated children being true positives, given general population prevalence (Y. S. Kim et al., 2011). Our attempts to use several machine learning approaches for our samples produced accuracy well below this. The impact of reducing the training set from 90% to 62% of samples, which resulted in considerably worse performance, suggests that the small number of control samples available was a significant factor. Therefore, we do not believe that our results invalidate attempts to build improved prediction models in future with larger training sets (Talebizadeh et al., 2008).

2.5.2. Post-transcriptional regulation

In this work, we present novel approaches to identify post-transcriptional regulation of microRNAs. Our algorithm for counting reads supporting isomiRs was restricted by limitations in quality of SOLiD sequencing reads at 3' ends, and by use of SeqTrimMap for alignment (Marco & Griffiths-Jones, 2012). We therefore performed our analysis of isomiRs distinguished only by 5' start site and strand, in contrast to other pipelines (de Oliveira et al., 2013). This approach was chosen to characterise some of the post-transcriptional regulation, although failing to capture the full detail of alternative isomiRs. Our methods could not detect non-templated nucleotide additions to either the 5' or 3' ends, as has previously been reported

(Bizuayehu et al., 2012). We performed hypothesis testing for difference in the proportion of transcripts from the dominant 5' isomiR against transcripts from other 5' isomiRs of the same mature microRNA. We used this method to amalgamate alternative 5' isomiRs to provide a simpler test across a large number of microRNAs with various patterns of isomiRs, including microRNAs with two known alternatives, and those with less consistent cleavage (Bizuayehu et al., 2012). Although it has previously been reported that the networks suppressed by isomiRs overlap with their reference microRNAs (Cloonan et al., 2011), it is not clear how much targeting properties differ between isomiRs. We calculated ratios for each sample and performed a Mann-Whitney U test, since this allowed us to handle cases with no reads for either form (recorded as zero or infinity or sample excluded if no reads for either form). Our approach fails to take into account difference in variance dependent on read depth, which might produce a more sensitive test (Anders & Huber, 2010). Although no microRNAs were identified as having differential usage of its 5' isomiRs after multiple testing correction, interestingly hsa-miR-146a-5p was inside the top 10 with p-value 0.0046 for difference in usage from an isomiR starting at chromosome 5, base 159,912,379 to starting at base 159,912,380. This same microRNA was identified as having significantly different arm usage when a similar Mann-Whitney U test was performed for read counts supporting the two arms of microRNAs. Although arm switching and differential isomiR usage have been reported previously, it is not clear what regulatory and processing mechanisms would drive this change. However, it has been shown that predicted target sets are radically different for alternative arms of microRNAs (Marco et al., 2010).

The microRNA hsa-miR-146a-5p is particularly interesting due to multiple facets of dysregulation identified, with possible isomiR differences as well as arm switching observed here and differential expression previously reported in multiple studies (Hicks & Middleton, 2016). Known SNP rs2910164 in the seed region of hsa-miR-146a-3p was notable for recurrence in the autistic group (found in 8 cases and 1 control). Previous studies suggested that this SNP might provide protection against digestive

cancer (W. Xu et al., 2011), although this was not found by subsequent meta-analysis (F. Wang, Sun, Zou, Fan, & Song, 2012). There is also a suggestion that the SNP may be associated with some characteristics of schizophrenia patients, although this is unclear (Zou et al., 2012). Differential expression of hsa-miR-146a-5p was reported in brain tissue (Mor et al., 2015), LCL (Talebizadeh et al., 2008) and olfactory stem cells previously (Nguyen et al., 2016). In an exploration of the possible functional implications of this microRNA, it was shown that autistic individuals had altered levels of some of the targets of hsa-miR-146a-5p (although the large number of targets in general means this could occur by chance), and that the microRNA is highly expressed in brain regions associated with higher cognitive functions in mice (Nguyen et al., 2016). Viral vector-mediated expression of the microRNA in mouse primary cell cultures caused changes to dendritic arborization in hippocampal neurons and astrocyte glutamate uptake capacities (Nguyen et al., 2016). This suggests that changes to hsa-miR-146a-5p expression could lead to functional impacts, although the network driving this change and the level of microRNA dysregulation required for significant effects is not clear. In our data, the abundance of the dominant 5p arm was not significantly different, and on average expression was reduced to 71% of control levels in autism. The arm switching effect that we observed was driven by this reduction combined with approximately quadrupled expression of the 3p arm. This test was affected by control samples in which no reads supported the 3p arm, which may reduce the confidence of the result. This discordant pattern of dysregulation in our samples suggests that functional consequences are unlikely to be similar to those previously explored. However, it suggests a potential for convergence of mechanisms of dysregulation on the microRNA. Further exploration of such mechanisms is required to help understand microRNA post-transcriptional dysregulation in disease, as well as how different convergent pathways might support autism susceptibility (Verhoeff, 2013).

Separate quantification of 5' isomiRs allowed us to perform differential expression analysis on this alternative dataset, an approach which may give

a more rich view of microRNA regulation (de Oliveira et al., 2013). We identified 24 5' isomiRs with significant differential expression in autism. Previous reports showed differential expression for hsa-miR-142-3p (Mor et al., 2015), hsa-miR-191-5p (Hicks et al., 2016; Sarachana et al., 2010) and hsa-miR-342 (Sarachana et al., 2010). The mechanism for hsa-miR-142-3p dysregulation may have been identified by the previous study, which found hypomethylation of the microRNA promoter in brain samples (Mor et al., 2015). Such a mechanism of general dysregulation in autism is supported by other work showing four other significantly differentially methylated regions in multiple brain regions in autism (Ladd-Acosta, Hansen, Briem, Kaufmann, & Feinberg, 2013). Although none of these regions overlap significantly differentially expressed microRNAs in our work, discovery of more fine-scale perturbation of DNA methylation in autism may explain multiple dysregulated microRNAs.

The differential expression in the region of mir-191 that we captured corresponded to the mature microRNA hsa-miR-191-5p. In a study of microRNA expression across brain regions for 18 normal human donor brains, this microRNA was identified as differentially expressed between the cerebellum and hippocampus (Ziats & Rennert, 2013).

Most importantly, our isomiR sensitive approach allowed recovery of a signal for differential expression in mir-23a. This corresponds to hsa-miR-23a-3p, which was repeatedly reported as differentially expressed previously (Abu-Elneel et al., 2008; Hicks et al., 2016; Talebizadeh et al., 2008; Y. E. Wu et al., 2016). Failure to identify this microRNA in the standard differential expression analysis may have been due to noise from poorly aligned overlapping reads not truly corresponding to the mature microRNA. Noise due to inconsistent processing of 5' ends for microRNAs may similarly explain the failure to find differential 5' isomiR usage between dominant and alternative for any single microRNA. This microRNA has the same seed sequence as the related hsa-miR-23b-3p, which has been implicated in autoimmune inflammation (Zhu et al., 2012). Interestingly, miR-23b-3p was close to the significance threshold for

differential expression in our dataset, at adjusted p=0.07. Regulation of inflammation is relevant in autism, as active neuroinflammatory processes have been reported in autism post-mortem brain samples (Vargas, Nascimbene, Krishnan, Zimmerman, & Pardo, 2005). Inflammation co-expressed genes have also been used to attempt to predict autism status in toddlers, although with limited success (Pramparo et al., 2015). We believe our work shows that regulatory microRNAs of such responses may be valuable additions to predictive models in future. The sequences for the two microRNAs 23a-3p and 23b-3p are highly similar, with only a single variation at base 19. Therefore, differentiating between expression levels for these two microRNAs was complicated significantly by issues with reads aligned to multiple places, as true reads from 23b-3p could be aligned at 23a-3p with a single mismatch. Excluding reads with multiple alignments is also a flawed approach as important microRNAs could not then be quantified.

2.5.3. Sample phenotypes

We used ADI-R interview scores across the IMGSAC dataset (IMGSAC, 2001), of which our samples are a subset, to investigate phenotypic clustering of individuals. We compared our results to those of a previous study which similarly identified three clusters and supported our use of the two first principal components through a factor mixed modelling analysis (Georgiades et al., 2013). Similar patterns of clustering and factors were also reported before, noting that such groups may reflect the structure of the ADI-R rather than true autism subtypes (Snow et al., 2009).

2.5.4. Implicated pathways for differentially expressed microRNAs

Among the 24 differentially expressed mature microRNAs in this study, several have previously been implicated in neurodevelopment or brain

functions. The strongest candidate for functional roles in autism may be hsa-miR-132-5p, which was reported as differentially expressed in autism previously (Abu-Elneel et al., 2008; Talebizadeh et al., 2008), and also reported as perturbed in schizophrenia and other disorders (Chan & Kocerha, 2012). Study of interactions with dendritic spine proteins showed that this microRNA was associated with Fragile X protein FMRP and it is also involved in activity-dependent regulation of neuronal morphology (Jin et al., 2004). This protein is well known for roles in neurodevelopmental disorders and syndromic autism (Crawley, 2012). Rat hippocampal neurons in culture that were transfected with hsa-miR-132-5p showed 'stubby and mushroom spines' at dendrites and increased average protusion width, with decreased protrusion density (Edbauer et al., 2010). These changes led to a strengthening of synaptic transmission in cultured hippocampal neurons, as measured by an increase in AMPA receptor-mediated miniature excitatory post-synaptic currents. The mechanism for miR-132-5p involvement appears to be through its association with FMRP, as knockdown of the gene reduces this effect (Edbauer et al., 2010). It is thought that microRNAs interact with FMRP indirectly, possibly through Argonaute proteins (Caudy, Myers, Hannon, & Hammond, 2002). The mechanism by which FMRP disruption leads to neurodevelopmental disorder is still not fully understood, however it is known that loss of FMRP leads to loss of translation repression for proteins that may remodel the cytoskeleton and internalise receptors, and other proteins involved in pathways for dendritic spine shape and synaptic plasticity (Bagni, Tassone, Neri, & Hagerman, 2012).

Alongside hsa-miR-132-5p, it was previously observed that hsa-miR-125b-5p had very high association with FMRP in the brain in mice (Edbauer et al., 2010). Our most significantly differentially expressed microRNA, hsa-let-7c-5p, was the fourth most highly associated microRNA to FMRP (Edbauer et al., 2010). Differentially expressed hsa-mir-138 was also associated with FMRP in mouse brain (Edbauer et al., 2010). However, our samples were all from multiplex families and did not include syndromic autism. The IMGSAC samples did not include any with Fragile X syndrome. This suggests that where microRNA dysregulation patterns

converge with Fragile X etiology, there is potential for more broad success from developing and applying Fragile X syndrome therapeutic approaches (Bagni et al., 2012).

The microRNA hsa-mir-138 was previously reported to be enriched in the brain and localised in dendrites, based on a screen for synaptosome microRNAs (Siegel et al., 2009). Transfection of rat hippocampal neurons with microRNA antagonists for hsa-miR-138-5p led to an increase in dendritic spine volume. This regulation was attributed to control of expression of APT1 through translation repression, with a binding site that when disrupted led to loss of control (Siegel et al., 2009). Within our data, we observed significant differential expression of only hsa-miR-138-1-3p, however the other arm hsa-miR-138-5p was close to the threshold with adjusted p=0.08. We also confirmed a prediction for targeting by this microRNA of APT1 (synonym LYPLA1) through miRanda screening, identifying a single binding site (Enright et al., 2003). We observed close to a doubling of expression of this microRNA in our samples, which may then lead to inhibited spine growth through the mechanism of APT1 affecting G protein alpha subunits previously reported (Siegel et al., 2009). Further exploration of the interaction of hsa-miR-138-5p with APT1 showed that the microRNA was correlated with mouse short-term memory, and that APT1 expression changes across age differed in mice with impaired memory (Tatro et al., 2013).

The second most significantly differentially expressed microRNA in our study, hsa-miR-181b-3p, was also previously reported as differentially expressed in autism (Ghahramani Seno et al., 2011). The important role of this microRNA in the brain is supported by differential expression of hsa-miR-181b-5p after stroke (Z. Peng et al., 2013). Experiments showing targeting of genes, and an antagomir reducing neural cell loss in cerebral ischemic cortex, as well as demonstration of mechanism through the genes HSPA5 and UCHL1 suggest that downregulation of hsa-miR-181b-5p can protect against injury after stroke (Z. Peng et al., 2013). Although we did not find the 5p arm differentially expressed in our sample, this microRNA was

one of 8 with significant change in arm usage in autism. This suggests that the microRNA may be subject to post-transcriptional regulation and complex mechanisms of control in the brain. Roles for this microRNA have also been proposed in schizophrenia, as a screen in postmortem cortical grey matter from the superior temporal gyrus showed differential expression of hsa-miR-181-5p (Beveridge et al., 2008).

Neurofibromatosis type 1 (NF1) is a tumor disposition syndrome that is strongly associated with autism (Garg et al., 2013). The differentially expressed microRNA hsa-let-7b-5p in our study was previously found to be dysregulated in NF1 (Masliah-Planchon et al., 2013). However, mechanisms proposed for the role of this microRNA in NF1 were largely related to pathways involved in tumorigenesis such as RAS-MAPK pathway regulation, mesenchymal transition and cell cycle progression.

Although we found significant differential expression of hsa-miR-32-5p, our result conflicted with opposite dysregulation reported previously (Hicks et al., 2016). It is possible that this could reflect different autistic individuals in the two samples having the same networks perturbed in different directions, leading to similar downstream effects. However, it should also be noted that different previous studies have identified a large number of alternative microRNAs which were differentially expressed in autism in different samples. In 12 studies comparing microRNA expression in autism cases and controls, a total of 185 microRNAs were found differentially expressed, with only 27 microRNAs showing consistent dysregulation in two studies, and 3 microRNAs consistent in three studies (Hicks & Middleton, 2016).

2.5.5. Chapter summary

Together, our results demonstrate significant dysregulation of microRNAs in autism. Convergence of findings for tests of post-transcriptional regulation of microRNAs demonstrates the value of improved methods for this under-studied area of microRNA regulation. Highlighted pathways implicated by

multiple microRNAs identified here and previously may lead to an improved understanding of gene networks affected in autism.

3. Target networks of variant microRNAs in autism identified through whole genome sequencing

3.1. Abstract

Autism is a neurodevelopmental disorder with a wide spectrum of phenotypes, involving difficulties with social interaction, communication and behaviour. Although autism is known to be highly heritable, individual genes and variants in general have only small effects on susceptibility. Instead, the genetic burden in autism risk appears to function through its convergence on shared pathways. Post-transcriptional regulators of gene expression such as microRNAs (microRNAs) have significant potential for affecting broad pathways in autism. As microRNAs are strongly conserved, rare variants are likely to have significant effects on large numbers of target genes. In this study, we report a survey of microRNA variation in autism from whole genome sequencing data covering 671 affected probands together with other family members. We identify 3949 single nucleotide variation (SNV) loci within affected individuals in annotated primary microRNA regions. We perform multiple filtering steps to identify candidate damaging variants, yielding 101 rare SNVs within microRNA seed regions. We generate a set of variant microRNA sequences based on these SNVs and predict variant target interactions. We show that rare variants are predicted to radically alter the target sets of affected microRNAs. We identify networks of genes targeted by these affected microRNAs and highlight possibly significant interactions affecting autism pathways.

3.2. Background

Autism is a neurodevelopmental disorder characterised by difficulties in social communication and interaction and restricted or repetitive behaviour

and interests (Lai et al., 2013). It is known that genetics plays a major role in autism susceptibility. Heritability has been estimated to be between 56% and 95% based on twin studies (Colvert et al., 2015). Studies based on polygenic scores for common variants give even stronger evidence that about 95% of variance in autism is accounted for by genetics (Gaugler et al., 2014). While much of the underlying susceptibility is due to small contributions by a large number of common variants, several genes are known to have a strong influence when affected by rare damaging variants. Some of these have been confirmed in animal models, for example where autism-like phenotypes have been observed in mice with Cntnap2 knockouts (Peñagarikano et al., 2011). It has been estimated that approximately 20% of autism cases can be explained by rare mutations affecting known autism genes (Berg & Geschwind, 2012). While the mechanisms of these causative variants are yet to be understood, it has been shown that affected genes are enriched in pathways regulating neuronal activity (Berg & Geschwind, 2012). Due to the substantial genetic heterogeneity of autism, it is likely that significant genes and pathways driving susceptibililty have yet to be identified.

MicroRNAs (microRNAs) are small RNA transcripts that work to regulate mRNA translation and abundance. They operate by base-pairing with binding sites within an mRNA 3' UTR (untranslated region). This leads to blocking of translation, or induces degradation of the mRNA. The binding of microRNAs to targets appears to depend largely on matching of sequence to the microRNA seed region, encompassing bases 2 to 8, but prediction models also include consideration of free energy and target site accessibility (Enright et al., 2003). microRNAs are thought to target over half of the genes in the human genome and have roles in most functional networks (Bartel, 2009).

microRNAs are remarkably well conserved, with relatively little variation between species and within humans (Friedman & Farh, 2009). This may be due to sensitivity in microRNA targeting to mutations, and also the large number of genes which each microRNA regulates. It has been shown that

microRNAs are good candidates in searches for rare variants affecting disease (Okada et al., 2016). Previous studies have suggested strong roles for specific microRNAs within neurodevelopment, including working in a timer system for differentiation of neural progenitor cells (Bonev et al., 2012). Other microRNAs are known to regulate genes involved in synaptogenesis and neurotransmission (Hara et al., 2017). Data suggests this may be achieved through controlling spatial localisation and compartmentalisation of translation to regions such as axons and synapses (Hicks & Middleton, 2016). microRNAs have been identified which are involved in neurodevelopmental disorders such as schizophrenia (Wright et al., 2013). Similar pathways are known to be affected in autism, and dysregulation in some cases may be due to dysregulated microRNAs targeting genes in these pathways. Approximately 70% of microRNAs are expressed in the central nervous system. Expression of the microRNAs can have distinctive patterns across child development and between different brain regions (Hicks & Middleton, 2016).

There have been several studies to date examining the expression levels of microRNAs in participants with autism and comparing to controls without autism diagnosis (Ghahramani Seno et al., 2011; Talebizadeh et al., 2008). All together, 12 relevant studies have identified 219 microRNAs with potentially differential expression in autism, 34 of which were in common among multiple studies (Hicks & Middleton, 2016). Several of these microRNAs are known to have differential expression in autism in the brain, for example with changes in expression between the superior temporal sulcus association cortex and the primary auditory cortex (Stamova, Ander, Barger, Sharp, & Schumann, 2015), or known to be involved in other neurodevelopmental disorders (B. Xu et al., 2010). Studies have attempted to identify shared regulated pathways of the targeted genes, however results are still inconclusive as to the role of the microRNAs in dysregulation in autism (Sarachana et al., 2010).

Regardless of understanding mechanics of regulation, microRNA expression has been proposed as a strong candidate for prediction of autism status.

Generally stable expression, and the good reflection of brain expression levels within peripheral blood, support this application of microRNA as a biomarker (Talebizadeh et al., 2008). Polygenic risk scores have also been developed for the prediction of autism status based on genetic markers (Gaugler et al., 2014). There is therefore good potential for using variants identified within microRNAs and combining these with known microRNA expression patterns to better predict autism status.

The MSSNG Project is a large-scale whole-genome sequencing endeavour, which aims to identify variants including intergenic single nucleotide variations (SNVs) and copy number variations (CNVs), as well as testing for rare SNVs affecting autism genes (Yuen et al., 2015). The project has performed whole-genome sequencing for over 2300 individuals to date. In this study, we analyse variant datasets downloaded from the MSSNG database (Yuen et al., 2015). We catalogue the variants observed within microRNAs and annotate these with comparison to parents and other data sources in order to identify those most likely to be involved in autism susceptibility. We investigate the predicted effects on targeting for candidate variants and explore changes to affected pathways.

3.3. Methods

3.3.1. Sample details

This study makes use of data generated by the MSSNG Project, including the Hospital for Sick Children (Toronto) and Autism Speaks Inc. Data was obtained from the MSSNG database release 22092015 (Yuen et al., 2016). After filters for quality, the sequencing data represents 671 probands, as well as associated family members. Diagnosis and phenotype fields were provided based on ADI-R (Autism Diagnostic Interview-Revised) followed by ADOS (Autism Diagnostic Observation Schedule) and clinical evaluation. The MSSNG Project obtained informed consent for all participants, as approved by the Research Ethics Boards at The Hospital for

Sick Children, McMaster University and Memorial Hospital (Yuen et al., 2016). Genomic DNA was extracted and sequenced using Illumina HiSeq2000 by the MSSNG project, using paired-end sequencing with insert size 500bp. Assessment of genome sequencing data quality by the MSSNG project indicated 96.8% coverage of the hg19 reference human genome, with average 56x sequence depth. No samples had below 95% coverage, and none had an average sequence depth below 50x (Yuen et al., 2015).

3.3.2. Data access and annotation

Our study recovered variant data for 1,871 miRBase v20 annotated primary microRNAs using the Google BigQuery Interface (Kozomara & Griffiths-Jones, 2011). We joined tables between sample phenotype information and variant data to provide fields for each candidate variant. The fields recovered and used in the subsequent analysis were: chromosome, position, reference allele, concatenated alternative bases, GATK SNV filter pass, minimum supporting reads for reference blocks (GATK-defined blocks called as having no variants), read counts supporting reference and alternate alleles, genotype quality, possible genotypes and genotype likelihood on Phred scale, sample ID, affection status, family ID, sex, and relation to family proband (Piskol, Ramaswami, & Li, 2013). Entries for 'reference blocks' were recorded separately, to be used as evidence that there was high confidence that no variants were present in parents for *de novo* variant calling.

3.3.3. Variant calling and *de novo* variant filters

Variants were categorised as 'passing' and annotated as such by the MSSNG Project based on GATK assessment using the genotype quality and likelihood fields (Piskol et al., 2013). We followed previously reported methods for the calling of *de novo* variants within the SNV dataset (Yuen et al., 2016). We used downloaded variant annotations for ForestDNM software (Michaelson et al., 2012) reporting the variant as *de novo* based on

comparison of the site with family trios. We combined this with a more direct examination of the reference block data. We required that variants marked as *de novo* in probands had reference blocks at respective positions in parents. We also performed filters as described to require that less than 70% of reads were reference, that under 5% of reads in parents were non-reference, that the site was not within a MSSNG indel, that Phred quality score was above 30, and that the site was not at a homopolymer run of 7 or more bases (Yuen et al., 2016).

3.3.4. Target prediction

Rare SNV locations were mapped to MiRBase annotated microRNAs and the reference fasta sequences were modified in a variant file to reflect the variants observed. The microRNA target prediction algorithm SeedVicious was used to predict targets with Ensembl release 75 human reference genes for these variant fasta sequences (Marco et al., 2012). We restricted this search to available 3' UTR sequences, with seven or eight matching bases required. We used the tool miEAA to assess functional enrichment of target genes (Backes et al., 2016). This tool uses target predictions for reference microRNAs from miRWalk, which is based on seed pairing in both 3' regions and other non-canonical binding sites (Dweep, Sticht, Pandey, & Gretz, 2011). Gene ontology classes are then assigned to microRNAs whenever a target gene is a member of a class, and enrichment analysis proceeds at the level of microRNAs.

3.3.5. Exploration of targets for variant microRNAs

To explore the consequences of variants within microRNAs on target sets, we generated modified microRNA reference files. We edited the reference fasta sequences from miRBase v20 for the mature microRNAs with variant rare seed alleles, taking into account microRNA strand (Kozomara &

Griffiths-Jones, 2011). We then performed target prediction using SeedVicious (Marco et al., 2012). This more basic method of prediction was used to reflect a simpler model of target binding by seed pairing. In practice, it is unclear to what extent target sets are transformed due to SNVs (Reczko et al., 2012).

3.3.6. Functional pathway exploration

Selected microRNAs and genes were advanced to IPA analysis in three groups – the set of microRNAs containing at least one rare seed-region variant; the set of autism risk genes with high numbers of predicted variant-altered targeting; the network of genes targeted by microRNAs affected and previously reported as dysregulated in autism. These three sets were input as pathways to IPA, and then the grow tool was used to link nodes by direct and experimentally confirmed known interactions. We performed a core analysis for enriched functions and pathways, and we performed filters for gene graphs using annotations for function to identify those likely involved in synaptic transmission or neural development.

3.4. Results

3.4.1. Variant annotation

We queried the MSSNG database dated 22/09/2015 using the Google BigQuery interface for evidence of SNVs in mature microRNAs and primary microRNA regions. A total of 8,062,974 variant records, and records of reference blocks, were retrieved. We used criteria for SNVs passing quality thresholds as described previously, requiring GATK passing (Piskol et al., 2013). We annotated subjects for autism status based on phenotype information from Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS) tests, and matched individuals to their family members. After filtering, our collection included

whole genome sequencing data for 671 probands with data for both parents, as well as 687 fathers, 694 mothers and 244 affected siblings.

3.4.2. Variant distribution

We made a total of 358,836 calls of SNVs and indels within primary microRNA regions in probands and affected siblings across the dataset. A total of 3949 unique variant positions were identified, of which 1930 were annotated in dbSNP build 138 and 851 were classed as common (defined as allele frequency > 1% in at least one dbSNP defined population) (Sherry et al., 2001). We mapped our SNVs to microRNAs and recorded those which were within mature sequences. We counted the number of SNV calls made across all subjects for each microRNA (Figure 3.1).

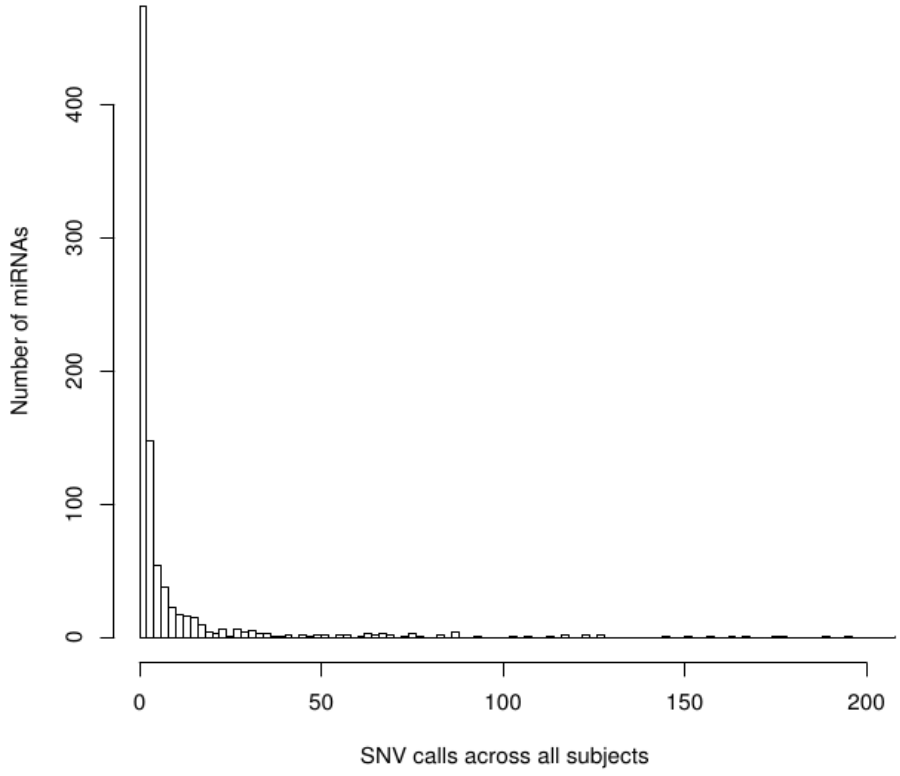


Figure 3.1. Number of SNVs detected in microRNAs

Number of SNV calls for each mature microRNA is calculated after filtering for quality. Calls are included for the same loci in a microRNA where they appear in multiple individuals, including probands, parents and siblings.

While it is known that common variants have a significant effect on autism susceptibility, microRNAs are strong candidates for rare variants having more severe individual effects (Berg & Geschwind, 2012). Therefore, we proceeded to investigate the rare variants within our dataset, defined as those which appeared at under 1% frequency in all dbSNP populations (Sherry et al., 2001). The locations of these variants are largely scattered across the genome, reflecting general positions of microRNAs through chromosomes (Figure 3.2).

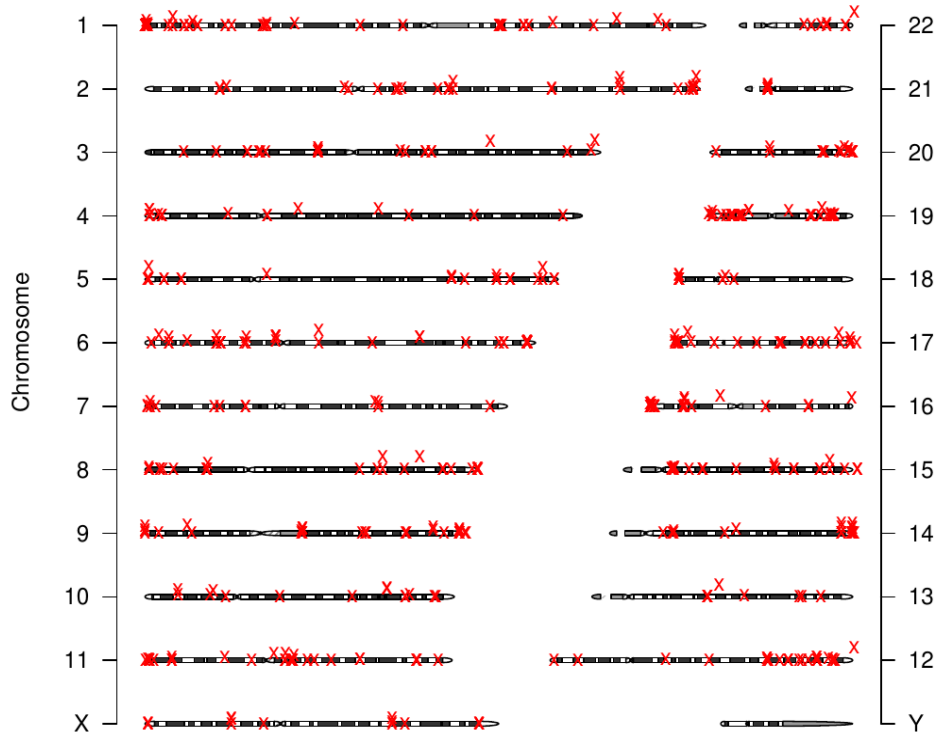


Figure 3.2: Rare SNV locations on chromosomes

Human chromosomes are displayed with cytobands painted by R package Quantsmooth. Rare SNV loci called in the dataset are identified with crosses. Cross elevation reflects relative frequency of SNV calls at a position across all samples.

We further used the conservation patterns of microRNAs and targeting models based on seed region matching to specifically focus on seed region variants, which would be expected to have the most powerful individual impact on large numbers of targeted genes (Lewis et al., 2003). To attempt to reduce false-positives due to systematic errors in alignment, we further required that each rare variant call should appear in no more than two families in the dataset. We noted that a very large number of the variant calls throughout our dataset appeared in very few individuals (Figure 3.3).

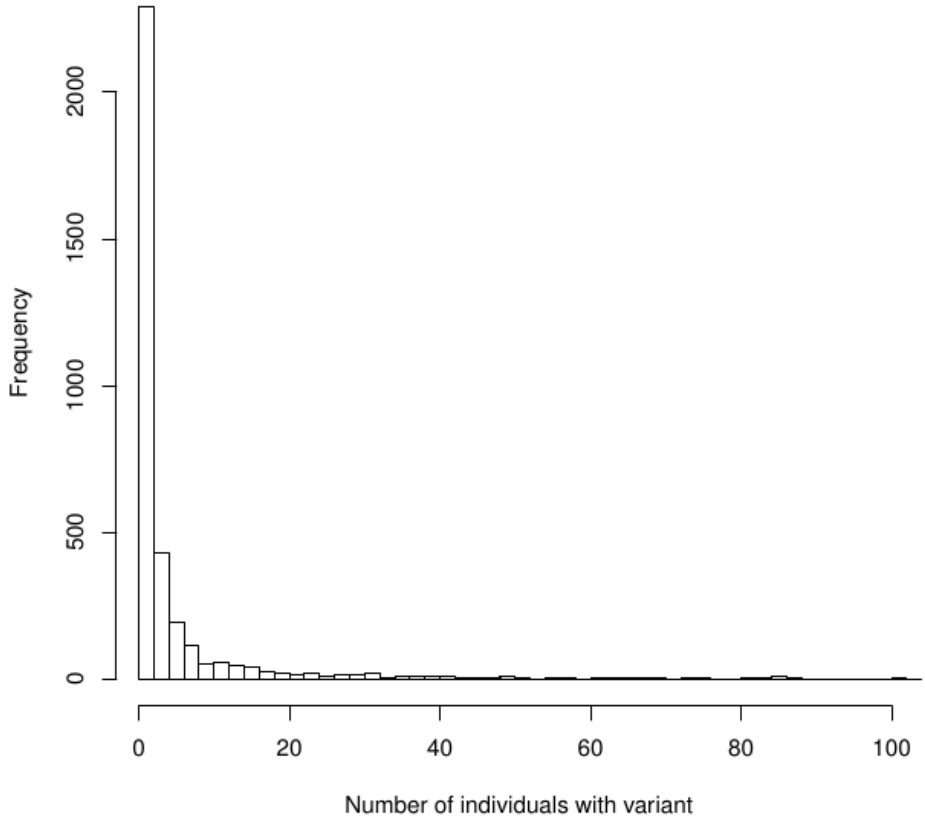


Figure 3.3. Number of individuals with each SNV

SNVs passing filtering criteria are included, and counted for their presence in probands, parents and siblings.

For variants which we filtered into the rare group, appearance in a large number of subjects suggests greater likelihood for a false positive. However, across all SNVs we do not observe a correlation between genotype Phred score and the number of individuals with a variant called, supporting the general robust quality of our filtered calls (Figure 3.4).

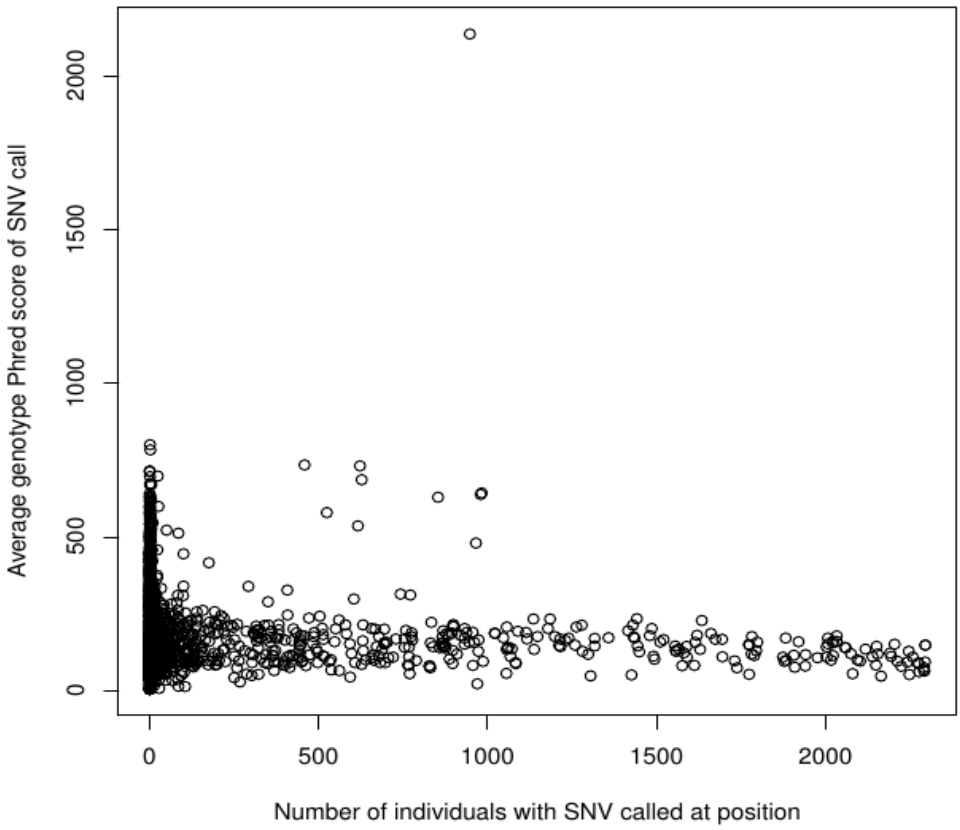


Figure 3.4. SNV call frequency versus genotype quality

For each SNV position, we calculate the average genotype Phred score of calls for that variant among all individuals carrying the variant. This is plotted against the total number of individuals in the sample in which the variant was called.

We identified 902 rare variant positions within mature microRNAs as defined by miRBase Release 20 (Kozomara & Griffiths-Jones, 2011). Of these, 275 positions were within microRNA seed regions. As we were interested in variants possibly affecting autism status, we filtered to only include those which were present in probands and affected siblings, and obtained 101 such rare, seed-region SNVs (Table 3.1).

Table 3.1. List of rare seed region variants

Chromosome	Position	Alternative Allele	Sample Sex	MicroRNA	dbSNP reference
1	17384	A	M	hsa-miR-6859-3p	rs201535981
1	30440	A	M	hsa-miR-1302	
1	10287790	G	M	hsa-miR-1273d	
1	37966554	C	M	hsa-miR-5581-3p	
1	60198922	T	M	hsa-miR-4711-3p	
2	25551537	G	M	hsa-miR-1301-3p	
2	69330864	A	M	hsa-miR-3126-3p	
2	80093691	C	M	hsa-miR-8080	
3	32547811	A	F	hsa-miR-548ay-3p	
3	48587443	T	M	hsa-miR-6823-5p	
3	49843617	A	M	hsa-miR-5193	
3	109321719	T	F	hsa-miR-4445-3p	rs543771609
3	122880650	T	M	hsa-miR-7110-5p	
3	195426331	A	F	hsa-miR-570-3p	
4	7461819	C	M	hsa-miR-4274	
4	7461822	C	M	hsa-miR-4274	

Chromosome	Position	Alternative Allele	Sample Sex	MicroRNA	dbSNP reference
4	172107344	T	M	hsa-miR-6082	
4	172107345	A	M	hsa-miR-6082	
5	32394552	T	F	hsa-miR-579-5p	rs548966059
5	71465309	G	M	hsa-miR-4803	
5	95414862	A	M	hsa-miR-583	rs555324098
5	143059451	G	M	hsa-miR-5197-5p	
5	154065391	C	M	hsa-miR-1303	rs370437195
6	31924714	T	M	hsa-miR-1236-5p	
6	33967804	T	M	hsa-miR-1275	
6	167411332	C	M	hsa-miR-3939	rs76608449
7	44921396	G	M	hsa-miR-4657	
7	134891776	A	M	hsa-miR-6509-3p	
8	1765417	T	M	hsa-miR-596	
8	92217763	G	M	hsa-miR-4661-3p	
8	120337431	G	M	hsa-miR-548az-5p	
8	125834245	C	M	hsa-miR-4662b	rs577795897
8	144895157	T	M	hsa-miR-937-3p	rs370660448
8	145486686	C	M	hsa-miR-7112-5p	rs7012034
9	68415337	C	M	hsa-miR-4477a	rs75019967
9	95290328	G	M	hsa-miR-4670-5p	
9	95290331	G	M	hsa-miR-4670-5p	rs539579265
9	96357156	A	F	hsa-miR-548au-3p	
9	97848320	G	M	hsa-miR-3074-3p	
9	140732887	T	F	hsa-miR-602	rs201175632
10	74480799	T	M	hsa-miR-	rs183259402

Chromosome	Position	Alternative Allele	Sample Sex	MicroRNA	dbSNP reference
				4676-5p	
10	98588456	G	M	hsa-miR-607	
11	209389	T	M	hsa-miR-6743-3p	rs377487596
11	61582657	T	M	hsa-miR-1908-3p	
11	111383675	G	M	hsa-miR-34b-5p	
11	128392326	A	M	hsa-miR-6090	
12	57912980	C	M	hsa-miR-616-3p	rs377525107
12	69666978	G	F	hsa-miR-1279	rs575587234
12	116866110	A	F	hsa-miR-4472	
12	116866114	A	F	hsa-miR-4472	
13	66792392	A	M	hsa-miR-4704-5p	rs532970894
13	90883501	G	M	hsa-miR-622	rs574914997
14	23426221	A	M	hsa-miR-4707-5p	
14	60113696	T	F	hsa-miR-5586-3p	rs549666512
14	65937871	G	M	hsa-miR-625-3p	
14	100575789	T	M	hsa-miR-151b	
14	100743704	A	M	hsa-miR-6764-5p	
14	101506576	T	F	hsa-miR-654-5p	rs533933068
14	106323701	A	M	hsa-miR-4539	
14	106324405	C	M	hsa-miR-4538	rs111207015
15	21038171	G	M	hsa-miR-3118	rs370569866
15	22049321	G	M	hsa-miR-3118	rs7167371
15	45725298	G	M	hsa-miR-147b	rs376183261
15	74703763	T	M	hsa-miR-6881-5p	
15	75646004	T	F	hsa-miR-631	

Chromosome	Position	Alternative Allele	Sample Sex	MicroRNA	dbSNP reference
15	85923874	T	M	hsa-miR-7706	
15	85923876	G	M	hsa-miR-7706	rs549230124
16	1785041	G	F	hsa-miR-3177-3p	rs555543222
16	2324668	A	M	hsa-miR-4717-3p	
17	29421432	C	M	hsa-miR-4733-5p	rs529859631
17	40666228	A	M	hsa-miR-5010-5p	
17	58120452	T	M	hsa-miR-4737	rs528487423
17	73494694	A	F	hsa-miR-6785-3p	
18	112282	C	M	hsa-miR-8078	
18	112283	G	M	hsa-miR-8078	
18	112285	G	M	hsa-miR-8078	rs9284388
18	112286	A	M	hsa-miR-8078	
19	1816187	T	M	hsa-miR-1909-3p	rs562400060
19	2234145	G	M	hsa-miR-1227-5p	rs112440628
19	2434927	A	M	hsa-miR-7108-3p	
19	2434992	A	M	hsa-miR-7108-5p	
19	10939658	A	M	hsa-miR-6793-5p	
19	13051335	T	F	hsa-miR-6515-3p	rs75453089
19	14640417	C	F	hsa-miR-639	rs377544418
19	47730217	T	M	hsa-miR-3190-5p	
19	54242644	A	M	hsa-miR-518a-3p	rs199969520
19	54251944	T	F	hsa-miR-521	
19	55899591	C	M	hsa-miR-6805-3p	rs368617558
20	1373566	T	M	hsa-miR-6869-3p	

Chromosome	Position	Alternative Allele	Sample Sex	MicroRNA	dbSNP reference
20	18451272	T	M	hsa-miR-3192-5p	
20	30194994	A	F	hsa-miR-3193	rs542196343
20	32636974	A	M	hsa-miR-4755-3p	
21	9825865	C	F	hsa-miR-3648	
21	15017121	C	M	hsa-miR-3118	
21	43166990	G	F	hsa-miR-6814-5p	
22	41649047	A	M	hsa-miR-6889-5p	
22	42296994	G	M	hsa-miR-33a-3p	rs77809319
22	46509625	G	M	hsa-let-7b-3p	
X	2527276	A	M	hsa-miR-6089	
X	49774344	G	M	hsa-miR-501-5p	
X	153285391	C	M	hsa-miR-718	

Variants are selected as rare seed region SNVs following the filtering process described. 37 out of 101 variants correspond to known dbSNP variants, although these are not classed as common, with no population above 1% frequency. Positions are given according to GRCh37 and are zero-based.

In order to further focus on those SNVs most likely to contribute to proband conditions, we finally filtered our SNV list to collect only those which were novel, and which had not been reported to dbSNP at all. This final collection contained 64 seed-region variants in probands and affected siblings (Figure 3.5).

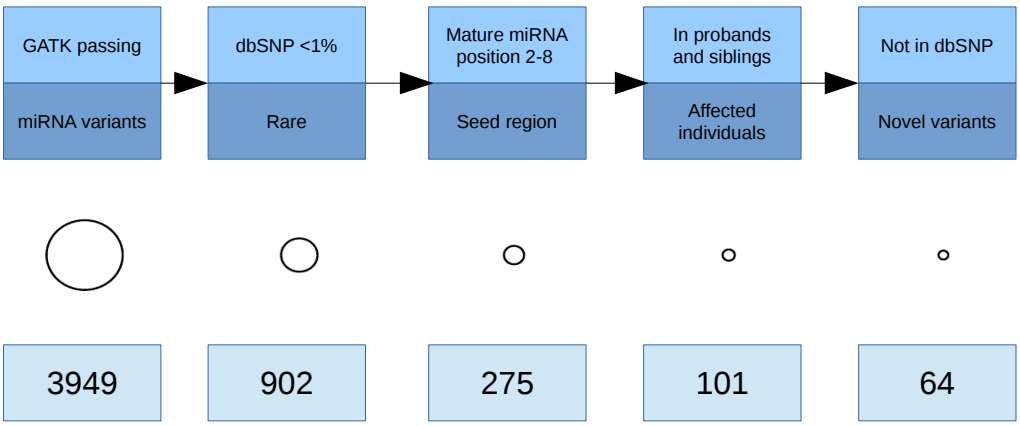


Figure 3.5. Flowchart of SNV filtering process

Five steps in variant filtering and prioritising are illustrated with the numbers of SNV loci remaining after each filtering step below. Circles are proportionate to the number of SNV positions in each group. In step one, GATK fields are used to select passing variants within microRNAs based on genotype quality. In the next filtering step, only those variants which are rare are selected, as determined by having less than 1% frequency in all dbSNP populations. Variants which are not recorded in dbSNP are also accepted. In the next filter, only variants within mature microRNA seed regions are used, as those most likely to affect targeting. We further filter to include only variants observed in subjects with autism diagnosis, excluding parents. Finally, the dbSNP filter is taken more strictly to require that a variant position is not reported in dbSNP.

3.4.3. Filters for *de novo* variants

As indications of potentially damaging variants, *de novo* mutations are strong candidates for drivers of autism susceptibility. However, there is a strong prior probability that *de novo* calls will be false positives, as they conflict with parental genotypes. Therefore we employed more strict filtering to attempt to identify such mutations. For candidate SNVs that were heterozygous in a proband or affected sibling, required that a reference block covered the region in both father and mother. We then implemented filters as described previously to further remove spurious candidates as described in Methods (Yuen et al., 2016). After these filters, none of the candidate *de novo* mutation sites remained. We calculated the expected number of *de novo* variants located within microRNAs based on the combined length of microRNA regions assessed and an approximate rate of *de novo* mutation of 1.20×10^{-8} per nucleotide per generation (Kong et al., 2012). The total length of microRNA regions examined was 153,114 bp, which suggests that we would expect 0.0018 microRNA *de novo* variants for each individual. Therefore, it is unsurprising that none were identified here, with an expected count of 1.23 for 671 autistic individuals studied.

Following a similar argument, that variants within seed regions are unexpected due to strong microRNA conservation, and that variants which are not common in dbSNP are less reliable calls, we investigated how robust our rare seed SNV calls were against these more stringent criteria, excluding those specific to *de novo* variant filtering based on comparisons between parents and probands. We confirmed that none of our SNV calls were within a homopolymer run of over 6 bases. Three out of 101 calls had a Phred score for genotype quality below 30. In another four out of the 101 calls, the proportion of reads supporting the reference allele was above 70%. Of the 64 entirely novel seed-region variants we report, 3 would fail more stringent criteria for *de novo* variation. Given these strong results, we considered the rare variant calls to be reliable.

3.4.4. Variant burden between groups

In aggregate, the genome-wide rare SNV burden in autism is not known to be higher than in controls. However, it has been reported that *de novo* SNVs which are predicted to be damaging to genes are more abundant sequencing data for autism cases (Yuen et al., 2016). Similarly, studies have reported a greater burden of predicted damaging variants in autism (T. Wang et al., 2016). While we did not have data for a control group in our sample, we were able to compare male probands against fathers without autism diagnosis. Although probands inherit their SNVs from their parents, the random pattern of recombination may lead to greater or lesser numbers of damaging mutations being inherited. We first tested for general SNV burden and found no significant difference. We then investigated whether autistic males and fathers had significantly different numbers of potentially damaging variants. Counting rare seed-region variants within the groups, we find 80 in fathers, 58 in male probands and 23 in male affected siblings. Given total numbers of samples in each group, with 687 fathers, 529 male probands and 192 male affected siblings, we find 0.12 variants per father and 0.11 variants per male autistic individual. Testing for a difference in proportions, we find no significant difference between the affected and unaffected groups. We restricted this test to males due to the relatively small number of female samples, and expected differences in SNV burdens between males and females (Bedford et al., 2016; Gilman et al., 2011). We investigated whether microRNA variants would follow the hypothesis that autistic females carry a greater burden of damaging variants than males (Pinto et al., 2014). We observed 20 such variants in 194 female autistic individuals, and 81 in 721 male autistic individuals, and found no significant difference in these proportions for females compared to males (0.10 versus 0.11 rare seed region variants per autistic individual).

3.4.5. Affected pathways

As microRNAs have not been well established previously as autism loci, we

searched for the presence of filtered rare variants within the set of 24 microRNAs which we identified as differentially expressed in a subset of IMGSAC samples. However, none were present within these mature microRNAs. Similarly, the MSSNG dataset had no rare variants within previously identified microRNAs hsa-miR-23a, hsa-miR-146a and hsa-miR-106b (Hicks & Middleton, 2016).

As a large number of microRNAs are affected by SNVs, an understanding of their functional significance requires exploration of pathways on which these microRNAs converge (Berg & Geschwind, 2012). We investigated the potential functional significance of our seed region variants through functional annotation. We used the miEAA tool which annotates microRNAs based on functions affected by their known targets and performs enrichment analysis at the microRNA level (Backes et al., 2016). We input the list of microRNAs which contained at least one rare seed-region variant in a proband or affected sibling. Testing for over-representation of pathways, gene ontology categories and diseases, we instead observed that 193 categories were significantly under-targeted.

We also tested the set of 101 rare seed-region variant affected microRNAs with gene set enrichment analysis (GSEA) using miEAA (Backes et al., 2016). We tested for enrichment for pathways, diseases (miRWalk-defined), gene ontology and diseases based on published studies on peripheral blood microRNA expression in the miEAA reference set (Backes et al., 2016). We found only a depletion (lower enrichment than expected by chance) for lung neoplasms ($p=0.03$).

In general, simply testing for functional enrichment of targets of microRNAs is not a good test for different functions between conditions, as we argue in the following chapter (Bleazard et al., 2015). However, to characterise the set of targets for the affected microRNAs in aggregate, we performed enrichment analysis of predicted targets using DAVID (D. W. Huang et al., 2009). A large number of functional classes were enriched (BP_FAT after multiple testing correction 175, KEGG pathways 59),

reflecting somewhat typical microRNA target sets.

3.4.6. Network analysis

We further explored the networks of affected microRNAs and their target genes through Ingenuity Pathway Analysis (IPA). The set of microRNAs containing at least one rare seed-region variant was input as a list and mapped to IPA records. We used the IPA knowledgebase to extend these microRNAs to a set of targets based on experimental confirmation, and alternatively to a set of targets using predictions, all with only direct interactions mapped. This produced a very large network (3943 nodes). We filtered the full network for genes annotated as involved in synaptic transmission or neural development. Performing an exploratory analysis of functions involved, we observed networks with known roles in neuronal cell adhesion. Enriched functions for the complete network included several with links to neuronal processes, however these were not necessarily the most significant, with a list including cancer and other processes.

3.4.7. Predicted effects of variants on microRNA targeting

For microRNAs, variant alleles are understood to have an impact primarily due to their effects on targeting of genes. We have focused on seed region variants in our analysis (Table 3.1) as these regions are strongly conserved and variants here are predicted to radically alter gene targeting (Enright et al., 2003). In order to assess the likely changes to microRNA targeting for our variant alleles, we ran the target prediction tool SeedVicious using modified mature sequences (Marco et al., 2012). On applying the SeedVicious algorithm, we predicted 856,026 microRNA-gene interactions for the unmodified microRNAs, and 862,364 target interactions for the mutated microRNAs.

We investigated how much the variant targets overlapped with the original targets of the microRNAs. We tested for whether the target sets were more similar to each other than expected by chance using Fisher's exact test. We found that there was no significant overlap in the two target sets (p=0.73). This indicates that the set of genes targeted by the mutated microRNAs was essentially no more related to the set of genes previously targeted than if selected randomly. Such a large change in targeting implies that functional shifts could be considerable. However, it is also possible that this change reflects only the nature of target prediction based on seed pairing, when in practice more complex access and free energy constraints lead to less alteration (Kertesz et al., 2007).

We used our set of predicted variant microRNA-target gene interactions to investigate targeting to known autism genes, as annotated by AutDB (Basu et al., 2009). We look at genes with particularly strong evidence for roles in autism through animal models and syndromic autism and convergence on neuronal regulation. *CHL1* and *CNTNAP2*, which function in neuronal cell adhesion and neurite outgrowth have 106 and 77 predicted microRNA interactions respectively (Peñagarikano et al., 2011). *FMR1*, a key gene involved in Fragile-X Syndrome, has 342 predicted interactions (Edbauer et al., 2010). *NLGN4*, involved in syndromic autism and also neuronal adhesion has 196 predicted interactions with variant microRNAs (Kenny et al., 2014).

3.5. Discussion

3.5.1. Assessment of dataset

Here we have reported a survey of microRNA variants in autism using whole genome sequencing data for 671 probands together with other family members. This dataset is the largest collection of whole genome sequences for autistic individuals to date, and provides an unprecedented opportunity to show variation in previously neglected regions (Yuen et al., 2015).

However, the number of probands investigated was still below thresholds recommended for successful statistical tests for implicated genes (L. Liu et al., 2013). Where previously studies on variants in autism focused on protein-coding genes (L. Liu et al., 2013), the sequencing data obtained has given us coverage over microRNAs, including primary transcript regions. This has revealed a very large number of microRNA variants present, with 3,949 unique variant positions. Studies using sequencing data for selected genes were previously biased to investigation of specific functional categories, such as synaptic function, however they confirmed an increased burden of rare damaging variants in autism cases (Kenny et al., 2014).

3.5.2. Abundance of microRNA variants

The majority of microRNA variants were common and present in many unaffected as well as affected individuals. However, a very striking number of rare variants were also observed. Of the 3949 unique positions at which variants were called, only 851 were common dbSNP annotated variants. The distribution of call counts shows that many of these loci had variants in only a single individual or a single family. A previous report on genome-wide mutations in a 200 proband subset of the MSSNG cohort found 9,774 germline *de novo* mutations (Yuen et al., 2016). As microRNAs are known to be well conserved in general, we predict that many of the variants we identified, and in particular the most rare seed-region variants, will have significant effects on gene expression regulation (Enright et al., 2003). The count of 101 rare seed region variants which we identified was at a similar order of magnitude to 31 loss of function variants found previously by targeted sequencing of synaptic function genes (Kenny et al., 2014). A larger targeted sequencing study of 2,446 autistic probands found 27 *de novo* variants. Better information about gene function and effects of mutations on protein sequence mean that direct associations between mutations and disorders are easier to establish with such data (O'Roak, Vives, Fu, & Egertson, 2012). Better development of microRNA annotations and knowledge may help to allow stronger inferences for rare variants within

microRNAs in future, for example through the Gene Ontology microRNA annotation project.

3.5.3. Tests do not find difference in microRNA variant burden

We tested for differences in rare SNV burden between male probands and fathers, and between male and female probands. However, no significant differences were identified. This result is in contrast to previous tests for categories of loss of function variants (Kenny et al., 2014) and CNVs (Pinto et al., 2010). Within the MSSNG dataset, a test was also performed previously for predicted damaging *de novo* mutations genome-wide, which found greater burden in autism cases (Yuen et al., 2016). This may be explained by difficulty in isolating those microRNAs and variants which impact autism risk from those which are incidental to the condition. Previous studies investigating the burden of deleterious alleles similarly failed to observe differences for aggregate variants, and only observed differences for variants selected for pathogenicity (Pinto et al., 2010). Further studies are required to differentiate microRNAs for their roles in autism and to identify classes of variants which are best for prediction of autism status.

3.5.4. Variant targeting properties

Discovery of microRNA variants allowed us to explore the predicted gene targets of the variant alleles. We found that the target sets of the microRNAs with rare seed-region variants were radically altered, such that they no longer resembled the original target sets. However, this result should be treated with caution due to limitations in target prediction algorithms (Ritchie et al., 2009). While pathway analysis on the aggregate microRNA-target networks was inconclusive, some interesting pairings in known autism networks emerged from analysis. This exploration was based on

previous information on known genes related to autism (Edbauer et al., 2010; Kenny et al., 2014; Peñagarikano et al., 2011).

3.5.5. Exploration of affected genes

Within our IPA analysis, we explored interactions between genes targeted by microRNAs with rare variants in our dataset. We found that a very large number of genes were predicted to be affected. Such large affected networks may be helpful to explain how small changes in genomes can create phenotypic changes in autism ranging from sensorimotor difficulties to language and behavioural changes, since a single microRNA may influence a similarly wide range of gene networks. However, actually discovering the specific pathways and causal relationships in such tangled and complex networks remains an unsolved problem.

We explored specifically the network involving *CNTNAP2*, *FMR1* and *NLGN4*. Target prediction through SeedVicious suggested 77 binding sites for *CNTNAP2*, 342 sites for *FMR1* and 196 predicted binding sites for variant microRNAs. Our IPA exploration of these genes showed that unfortunately interactions are still poorly understood for *CNTNAP2* and *NLGN4*, despite these genes having known causative roles in autism cases (Peñagarikano et al., 2011). However, the downstream genes affected by *FMR1* dysregulation have been widely studied (Carter & Scherer, 2013). We suggest that interactions demonstrated, for example with three binding sites for the variant version of hsa-miR-6082, could potentially play a role in the autism phenotype for some subjects.

4. Bias in microRNA functional enrichment analysis

4.1. Abstract

Many studies have investigated the differential expression of microRNAs in disease states and between different treatments, tissues, and developmental stages. Given a list of perturbed microRNAs, it is common to predict the shared pathways on which they act. The standard test for functional enrichment typically yields dozens of significantly enriched functional categories, many of which appear frequently in the analysis of apparently unrelated diseases and conditions. We show that the most commonly used functional enrichment test is inappropriate for the analysis of sets of genes targeted by microRNAs. The hypergeometric distribution used by the standard method consistently results in significant p-values for functional enrichment for targets of randomly selected microRNAs, reflecting an underlying bias in the predicted gene targets of microRNAs as a whole. We developed an algorithm to measure enrichment using an empirical sampling approach, and applied this in a reanalysis of the gene ontology classes of targets of microRNA lists from 45 published studies. The vast majority of the microRNA target sets were not significantly enriched in any functional category after correction for bias. We therefore argue against continued use of the standard functional enrichment method for microRNA targets.

4.2. Background

MicroRNAs (microRNAs) down-regulate abundance and translation of target mRNAs through complementary binding to target sites. microRNAs play important roles in regulating gene expression in response to stimuli and during development and their expression patterns can be predictive of disease states (Leidinger et al., 2013; Schratt et al., 2006; Xie et al., 2013). For this reason, a large number of studies have investigated the expression

of microRNAs in a wide range of biological conditions. Microarray assays, qRT-PCR and high-throughput sequencing have all been used to identify differentially expressed microRNAs in disease states, between different tissues and during development (Davidson et al., 2010; Liang, Liu, & Ye, 2013; C. Wu et al., 2012). Unfortunately, the interpretation of microRNA differential expression is not straightforward. The roles of individual microRNAs in cellular pathways are still poorly understood. Each microRNA has the potential to target hundreds of different genes, meaning that perturbation of a single microRNA may affect many biological functions (Friedman & Farh, 2009). This motivates a broad view: given a list of differentially expressed microRNAs, we must look for the functions or pathways on which they converge.

Here, we examine the most common method of microRNA functional enrichment analysis, used in over 45 published studies. This method consists of three steps: finding which genes are targeted by selected microRNAs; annotating target genes for their participation in pathways and processes; and statistical testing for over-representation of a biological process in the set of targeted genes (Gusev, Schmittgen, Lerner, Postier, & Brackett, 2007). For the first step, computational target prediction is usually necessary because experimental datasets covering microRNA-mRNA interactions on a genome scale are currently lacking. For the second step, annotation by GO term membership or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways is common (Ashburner, Ball, Blake, & Botstein, 2000; Kanehisa & Goto, 2000). For the final step, the hypergeometric distribution, or equivalently Fisher's exact test, is used to test for enrichment. The hypergeometric distribution describes the situation where samples are picked uniformly at random from a finite population which contains a labelled subset. In the context of functional enrichment, it gives the probability of targeting k genes from a labelled category when targeting a total of n genes from the genome. We can then use this distribution to test the null hypothesis that genes were targeted randomly versus the alternative that genes belonging to a given annotation were preferentially targeted.

This approach, which we here refer to as the 'standard method' because of its preponderance, consistently produces a large number of significantly enriched processes. However, these are often difficult to interpret, and full lists of significant terms are rarely provided in published articles. Recurrence of GO terms between apparently unrelated diseases and conditions in the literature is very notable (T.-P. Lu et al., 2012). Most worryingly, even random and meaningless microRNA lists produce significant functional enrichments using the standard method (Ritchie et al., 2009).

In this study, we show that critical problems with the standard method arise because of bias in the sets of genes that are predicted to be targeted by microRNAs in general. This means that the assumption of uniform sampling in the hypergeometric distribution is not reasonable. We correct for this bias by bringing the statistical test back from the level of genes to the level of microRNAs, and show that most functions reported as significantly enriched in the literature do not remain so after correction.

4.3. Methods

We developed an algorithm to find the empirical distribution of the number of microRNA target genes within annotated functional categories. We predicted targets of all miRBase release 20 annotated mature microRNAs in the 3' UTRs of all Ensembl release 75 human and mouse genes using miRanda (version 3.3a, free energy < -20 kcal/mol, score > 155) (Enright et al., 2003). Annotated biological process GO terms for all human and mouse genes were downloaded from Ensembl (Ashburner et al., 2000; Kinsella et al., 2011). Following the standard method, we defined the target genes for a list of microRNAs as the union of genes predicted to be targeted by each microRNA. We then calculated GO term overlap as the proportion of target genes which were annotated as belonging to a given GO term. Our empirical algorithm first counted the GO term overlap for targets of an input microRNA list. A set of microRNAs of the same size as the input list was

randomly sampled without replacement from the set of all annotated microRNAs, with one million iterations. An empirical p-value for each GO term was calculated using the proportion of simulations that produced an equal or greater GO term overlap. We also developed a modified multihit version of this algorithm that did not use the simple union of target genes, but instead gave each gene a score given by the sum of the number of predicted target binding sites for all input microRNAs. GO term overlap was then defined as the score for genes annotated with a given GO term divided by total score for input microRNAs. We repeated all analyses using KEGG pathways (accessed through the KEGG.db Bioconductor package, which archives KEGG data from March 15th 2011 (Kanehisa & Goto, 2000)) in place of GO terms. We also recalculated target predictions based on the intersection of genes returned by three alternative target prediction algorithms: PITA predictions from the PITA Targets Catalog version 6 (August 2008) based on mm9 and hg18 for mouse and human respectively, with zero flank and all sites included (Kertesz et al., 2007); DIANA-microT-v4 predictions, which use miRBase annotated microRNAs and Ensembl 3' UTR sequences (Reczko et al., 2012); and canonical seed matches between miRBase microRNAs and Ensembl 3' UTR sequences as in (Bartel, 2009). We used miRBase alias data to match the names of microRNAs in downloaded prediction sets to their current annotations (Kozomara & Griffiths-Jones, 2011), and Ensembl gene synonyms to match gene names in the target sets to Ensembl GO classifications. A further alternative was included to reflect the approach of miRFunction in combining computational target prediction with CLIP-seq experimental datasets(J.-H. Li et al., 2014). 36 human and 5 mouse AGO pulldown CLIP-seq datasets were downloaded from starBase(J.-H. Li et al., 2014). Target predictions from miRanda were mapped to genomic loci and filtered to include only those supported by at least one experimental dataset. We also investigated the use of the standard method in studies of plant microRNAs. Following a typical analysis, we predicted TIGR genome cDNA (OSA1R5) targets of miRBase annotated rice microRNAs using psRNATarget with default parameters(Dai & Zhao, 2011). We matched the TIGR loci to gene ontology annotations in tables downloaded from the agriGO database(Du, Zhou,

Ling, Zhang, & Su, 2010). We performed a literature survey to identify studies that followed the standard method. We used the search functions provided by Nature Publishing Group and Public Library of Science, as well as Google Scholar, with the search terms 'gene ontology' and 'microRNA' for mouse and human, and 'oryza' and 'microRNA' for rice. Each article was manually checked to confirm that the standard method was followed. Lists of microRNAs reported to be differentially expressed (or otherwise flagged) were manually compiled from the retrieved articles. Where multiple lists of microRNAs were assessed in the same manuscript, we arbitrarily chose one list for testing. We define results as significant by default where $\alpha < 0.05$. In the case of multiple testing, we perform Benjamini-Hochberg adjustment and report significant items passing the threshold for false discovery rate < 0.05 (Benjamini & Hochberg, 1995).

4.4. Results

4.4.1. Assessing the appropriateness of the hypergeometric distribution

We used our algorithm to investigate whether the null hypothesis used by the standard method was appropriate by comparing the hypergeometric distribution with an empirical distribution for the number of predicted target genes belonging to a GO term for randomly sampled microRNAs. As an illustration, we use the GO term ‘ion transport’ (GO:0006811), which is often reported as significantly enriched in the literature (G. Liu et al., 2010; Sokolov, Panyutin, & Neumann, 2012; Yunta et al., 2012). We predicted the targets of an example set of 39 microRNAs that were reported as differentially expressed in one study (Sokolov et al., 2012). These microRNAs were predicted to target 10,057 genes out of 15,733 genes with at least one assigned biological process GO term, of which 327 are annotated for ion transport. These parameters were chosen to mirror the methods used by the popular tool DAVID (D. W. Huang et al., 2009). The

expected distribution of the number of target genes assigned to the ‘ion transport’ GO term according to the hypergeometric distribution is shown in Figure 4.1, alongside the distribution of the number of target genes of 1 million randomly chosen sets of 39 microRNAs. The data clearly show that the hypergeometric distribution does not adequately model the empirical background distribution under these conditions.

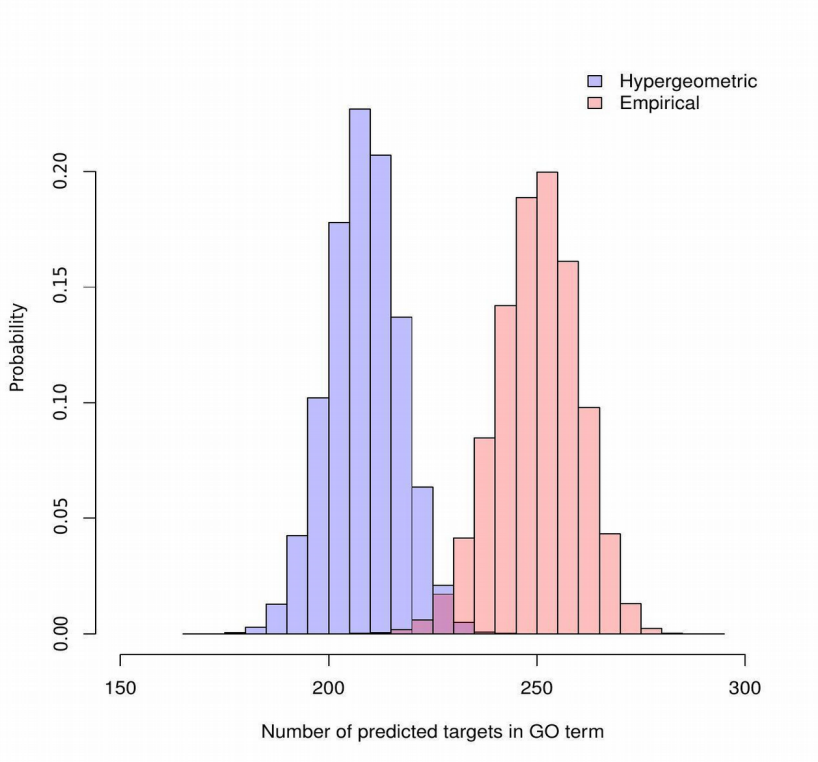


Figure 4.1. Expected and empirical number of predicted targets of randomly selected microRNAs in ion transport

For an example 39 microRNAs, we calculate the hypergeometric distribution (blue) for the number of expected targets in the GO term 'ion transport' (GO:0006811). The empirical distribution (red) represents the predicted targets of random samples of 39 microRNAs. The probability for each 5-gene bin is given according to both distributions.

Figure 4.1 immediately suggests an explanation for the excess of significant GO terms under the standard method. A typical microRNA target gene set, with GO membership near the mean of the empirical distribution, will produce significant p-values for GO term enrichment using the standard method. Indeed, the mean number of targets involved in ion transport for 39 random microRNAs (250 genes) gave a p-value of 5.97×10^{-7} when tested using the hypergeometric distribution. We compared the hypergeometric with the empirical distributions for all biological process GO terms. For each GO term, we constructed the hypergeometric distribution using the same example of 39 microRNAs targeting 10,057 genes as predicted by miRanda, again mirroring a common DAVID analysis (Enright et al., 2003; D. W. Huang et al., 2009). For each GO term, we then generated an empirical distribution for the number of member genes targeted by 39 randomly selected microRNAs. We calculated the one-sided p-value on the hypergeometric distribution of the mean of the empirical distribution for each GO term. Among the terms yielding the smallest p-values (Table 4.1), several are notably often reported as enriched in the literature, such as 'regulation of transcription, DNA-dependent' (GO:0006355) (Kraemer, Chen, Henning, & Faust, 2013; Mizuguchi et al., 2011; Munch, Harris, & Mohammad, 2013; Ziats & Rennert, 2013).

Table 4.1. GO terms with the largest difference between hypergeometric and empirical background distributions

GO term	Hypergeometric p-value of empirical mean
GO:0006351~transcription, DNA-templated	1.21×10^{-28}
GO:0006355~regulation of transcription, DNA-dependent	6.99×10^{-25}
GO:0007165~signal transduction	4.63×10^{-18}
GO:0006468~protein phosphorylation	4.34×10^{-17}
GO:0055085~transmembrane transport	1.39×10^{-13}
GO:0015031~protein transport	1.95×10^{-13}
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	3.44×10^{-12}
GO:0045893~positive regulation of transcription, DNA-dependent	4.80×10^{-12}
GO:0048011~neurotrophin TRK receptor signaling pathway	7.67×10^{-12}
GO:0007264~small GTPase mediated signal transduction	1.22×10^{-11}

Using 39 microRNAs targeting 10,057 genes, we calculated the p-value on the hypergeometric distribution for the rounded mean of the empirical distribution for each GO term. The 10 processes with the most extreme bias are shown.

Other GO terms suffered from an opposite bias, making detection of a significant enrichment almost impossible under the standard method. The most extreme of these with more than 5 members were 'defense response to bacterium', 'detection of chemical stimulus involved in sensory perception of smell' and 'G-protein coupled receptor signaling pathway', with p-values close to 1. These terms were in turn conspicuously absent from published lists of enriched processes in disease. These results imply that using the hypergeometric distribution to model microRNA target gene GO term membership is inappropriate and is liable to produce spurious results.

4.4.2. Re-analysis of published microRNA lists

We investigated the prevalence of the standard method with a non-exhaustive manual search of journal articles. We identified 40 published studies where the standard method was applied to investigate functional enrichment of targets of sets of microRNAs (Supplementary Table 4.1). Among these, a wide range of algorithms and their combinations were used for target prediction. Six different organisms were represented in studies ranging from sea cucumber aestivation to pseudorabies virus in pig cell lines (M. Chen, Zhang, Liu, & Storey, 2013; Y.-Q. Wu et al., 2012). A large number of different web servers provided tools for functional annotation and statistical testing of enrichment. The types of functional categories tested included biological process GO terms with various filters and KEGG pathways (D. W. Huang et al., 2009). The list included recent and high-impact publications (Soh, Iqbal, & Queiroz, 2013). Although functional enrichment analysis was generally not the central focus of these studies, it was mentioned in 26 out of 40 abstracts. Only a small subset of the published studies we surveyed provided a full list of significantly enriched functional categories. Several studies, however, reported in the main manuscript on a few significant GO terms appealing for interpretation (Ma et al., 2013). This may be misleading for readers who are unaware that hundreds of other GO terms are equally enriched. Although not fundamental to the standard method, this problem is exacerbated by its unfailing

production of large numbers of significant results.

We were able to collect lists of differentially expressed (or otherwise flagged) microRNAs from 22 studies in humans and 7 studies in mice for analysis with our empirical algorithm. Where a study performed enrichment analysis for multiple microRNA lists, we arbitrarily selected one list per published study. In order to mirror the approaches in published studies as closely as possible, where each microRNA was analysed separately, we also applied the empirical algorithm in a single test of only one microRNA. We converted microRNA names to their current annotations, removing 10 microRNAs whose miRBase entries had been deleted since publication (Kozomara & Griffiths-Jones, 2011). For each input microRNA list, we ran our algorithm with miRanda predictions (Enright et al., 2003), biological process GO term annotations (Ashburner et al., 2000; D. W. Huang et al., 2009) and one million iterations of randomly selected microRNAs, generating empirical distributions of GO term target gene overlap for the specified numbers of input microRNAs, and outputting p-values for the enrichment of GO terms.

After Benjamini-Hochberg correction for multiple testing (Benjamini & Hochberg, 1995), we observed an enrichment of any GO term in only 5 out of 22 human and 0 out of 7 mouse studies (Table 4.2). In contrast, all of the published studies reported multiple enriched functional categories. Although our aim is to provide a controlled comparison of standard and empirical methods, rather than to attempt to replicate the exact prediction and annotation methods of the previous studies, these results show that most functional categories reported to be enriched in the literature would not remain so after correction for bias.

Table 4.2. Results of empirical algorithm applied to published microRNA lists

Reference	Species	MicroRNA List Extracted	Significant GO Terms with Basic Empirical Algorithm	Significant GO Terms with Multi-hit Empirical Algorithm
(Arndt et al., 2009)	Human	1	0	0
(Chartoumpekis et al., 2012)	Mouse	10	0	17
(G.-Q. Chen, Zhao, Zhou, Liu, & Yang, 2010)	Human	176	6017	1193
(W.-S. Chen et al., 2013)	Human	15	0	6
(Cheng, Pastrana, Tavazoie, & Doetsch, 2009)	Mouse	1	0	0
(Collino et al., 2010)	Human	11	0	0
(Davidson et al., 2010)	Human	1	0	0
(Flavin et al., 2009)	Human	2	0	0
(He et al., 2013)	Mouse	8	0	0
(Hunter et al., 2008)	Human	1	0	0
(J. Jiang et al., 2008)	Human	18	0	0
(Keck-Wherley et al., 2011)	Mouse	12	0	0
(Kraemer et al., 2013)	Human	9	0	0
(Liang et al., 2013)	Mouse	37	0	0
(G. Liu et al., 2010)	Mouse	27	0	1
(Mizuguchi et al., 2011)	Human	3	0	0
(Munch et al., 2013)	Human	10	0	22
(Presneau & Eskandarpour, 2013)	Human	5	0	0
(Raponi, Dossey, Jatkoe, Wu, & Chen, 2009)	Human	15	0	239
(Romero-Cordoba et al., 2012)	Human	130	5573	206
(Sanchez-Diaz, Hsiao,	Human	26	2502	28

Reference	Species	MicroRNA List Extracted	Significant GO Terms with Basic Empirical Algorithm	Significant GO Terms with Multi-hit Empirical Algorithm
& Chang, 2013)				
(Schonrock et al., 2010)	Mouse	1	0	0
(Soh et al., 2013)	Human	1	0	0
(Sokolov et al., 2012)	Human	39	0	337
(Tanic & Andrés, 2013)	Human	46	3353	317
(X. Wu et al., 2011)	Human	25	2480	138
(Z. Yan et al., 2012)	Human	1	0	0
(Zhang, Daucher, Armistead, Russell, & Kottilil, 2013)	Human	1	0	0
(M N Ziats & Rennert, 2013)	Human	25	0	0

We investigated the use of the standard method in plant studies, and found 5 published articles where it was applied in the analysis of rice microRNAs. Four of the microRNA lists extracted from these studies could be input into our algorithm. The lists we extracted contained larger numbers of microRNAs than in other species, perhaps reflecting a preference of authors for simpler analyses of smaller and more manageable plant microRNA target sets. We ran our algorithm with these input lists using psRNATarget predictions and gene ontology annotations obtained from agriGO (Dai & Zhao, 2011; Du et al., 2010). We found significant enrichments in 2 out of the 4 input lists (Table 4.3), suggesting a similar general pattern to that in humans and mice.

Table 4.3. Results of empirical algorithm applied to published microRNA lists in rice studies

Reference	Species	MicroRNA List Extracted	Significant GO Terms with Basic Empirical Algorithm
(Abrouk et al., 2012)	Rice	69	34
(T. Peng et al., 2011)	Rice	90	0
(Tong et al., 2013)	Rice	153	Not tested
(Wei, Yan, & Wang, 2011)	Rice	68	0
(Yi et al., 2013)	Rice	142	49

4.4.3. Testing for multi-hit convergence on processes

Our results are robust to changes within the general framework of the standard method. Subsets of GO are often used in the literature. We therefore repeated all analyses using the filtered GO term annotation set BP_FAT downloaded from the DAVID Knowledgebase (D. W. Huang et al., 2009), with very similar results. As an alternative to biological process GO terms, we also used KEGG pathway annotations. Running the empirical algorithm with the published microRNA lists, 8 out of 29 microRNA lists produced at least one significantly enriched KEGG pathway. It is common to predict targets of microRNAs using several programs and use the intersection set of their outputs. As an alternative to prediction by miRanda (Enright et al., 2003) alone, we used the intersection set of target predictions by PITA (Kertesz et al., 2007), DIANA-microT-v4 (Reczko et al., 2012) and seed matching using canonical seeds (Bartel, 2009). These downloaded prediction sets did not include all currently annotated microRNAs; in particular 134 of the microRNAs from input lists were missing and so had to be excluded. Using this prediction method, 3 of the 29 microRNA lists from

published studies produced significantly enriched biological process GO terms. We also tested using a filtered target prediction set that only included miRanda target loci found by CLIP-seq experiments as in miRFunction (J.-H. Li et al., 2014). This produced similar results, with no significant enrichments for 18 out of 22 human and 7 out of 7 mouse microRNA lists.

We tested whether our algorithm was able to detect functional enrichment when the input microRNAs were artificially selected for their targeting of a given process. We manually selected as an input set the 8 microRNAs with the most predicted targets in the process 'regulation of axonogenesis' (GO:0050770). As expected, the algorithm found that the same GO term was significantly enriched, as well as other related and unrelated terms.

The standard method in its simplest form counts each gene once, whether it is targeted by one or many differentially expressed microRNAs, losing key information on functional convergence (Gusev, 2008; S. Y. Lee et al., 2012). Filters on target gene sets or on output GO terms have been proposed previously to resolve this problem (Gusev, 2008; Z. Hu et al., 2014). These filters require the proportion of microRNAs targeting a gene and the proportion of microRNAs with at least one target in a GO term to pass defined thresholds. Another alternative is to apply the statistical test on microRNA-mRNA pair connections (S. Y. Lee et al., 2012). We therefore modified our algorithm to address this issue while maintaining our simple hypothesis testing approach and the principled comparison with the empirical distribution. The set of target genes for an input microRNA list was previously defined as the union of predicted target genes. In our modified algorithm, each gene is assigned a score for strength of interaction with microRNAs based on the total number of predicted binding sites, including multiple sites for the same microRNA. The score for a GO term is then the sum of gene scores for its members, divided by the total number of binding sites for the microRNAs. As above, we run one million iterations with randomly sampled microRNAs and compare the GO term scores for differentially expressed microRNAs with this empirical distribution.

Results from our modified algorithm applied to previously identified microRNA lists are shown in Table 4.2. Significant enrichments were found for all the input lists which had positive results for the original basic algorithm, albeit with more modest numbers of significant GO terms. Enrichments were also detected from six lists where previously they were not found.

4.5. Discussion

Our comparison of the hypergeometric and empirical distributions showed that certain functional categories are preferentially targeted by microRNAs, regardless of whether those microRNAs are differentially expressed in a biological state or not. It is not helpful to report a GO term as enriched for targets of differentially expressed microRNAs if an equally strong enrichment would be obtained for randomly picked microRNAs. This justifies an empirical sampling approach, which measures enrichment relative to other microRNAs, in comparison to the standard method which does not take into account the background level of targeting of a GO term. There are several possible explanations for the phenomenon of preferential targeting by microRNAs, including bias in target prediction algorithms, similarities among seed sequences, correlations between genes that are regulated together, and genuine preference for control of certain biological processes by different mechanisms. One clear source of bias is average 3' UTR length of genes annotated to specific GO terms. When we use the p-value of the mean of the empirical distribution on the hypergeometric distribution (Table 4.1) as a measure for bias of a GO term, we observe a strong negative correlation with average 3' UTR length of genes assigned to that GO term (Pearson's $r = -0.36$; $p = 5.3 \times 10^{-287}$). We also note that many GO terms were invariably returned together as enriched. This may reflect underlying correlations between targeting of processes as well as the hierarchical structure of the gene ontology.

The simplest use of functional enrichment tests examines a set of genes with

a common characteristic – for example, a set of differentially expressed genes, or a set of genes with particular genomic properties. The test described here is subtly different: a set of microRNAs is defined by differential expression, and that set is one step removed (by microRNA target prediction) from the set of genes whose functional enrichment is tested. The bias in the underlying expected distribution of functional categories comes from the process of linking microRNAs with their target genes. While we have examined only a specific use of the functional enrichment test, similar biases may affect other genomic enrichment tests (Slowikowski, Hu, & Raychaudhuri, 2014). For example, CHIP-seq identification of transcription factor binding sites followed by functional enrichment of the target gene set is analogous to the analyses described here. Further investigation is therefore required to determine the appropriateness of the hypergeometric distribution for other types of functional enrichment studies.

In our literature survey, we identified 45 journal articles that used the standard method. However, our list was not exhaustive, and excluded studies with relatively minor deviations from the standard method, such as those that implemented more complex filters for selection of target genes (Cho et al., 2013) and those that performed tests based on combined targeting by microRNAs (S. Y. Lee et al., 2012). Altogether, we estimate that hundreds of published articles are likely to be affected by the bias described here. In the application of the empirical algorithm to microRNA lists from these studies, we do not attempt to directly replicate methods used, and instead show a pattern in results that strongly suggests that most reported enrichments would not be found with correction for bias. We also do not directly compare the significant enrichments output by our algorithm with those reported, as the specific types of functional categories assayed, such as GO term collections assembled by DAVID (D. W. Huang et al., 2009), vary greatly, and full lists of significant terms are rarely published.

Our results do not imply that differentially expressed microRNAs do not converge on functions of interest. It is notable that the number of

significantly enriched GO terms increased with the number of microRNAs input to our algorithm. This may be because noise dominates for smaller lists, whereas a larger number of input microRNAs provides more signal of convergence on a process. While modest enrichment of a function for a single microRNA is undetectable, the combination of many small enrichments for a larger collection of microRNAs passes significance thresholds. Five of the largest input microRNA lists (G.-Q. Chen, Zhao, Zhou, Liu, & Yang, 2010; Romero-Cordoba et al., 2012; Sanchez-Diaz, Hsiao, & Chang, 2013; Tanic & Andrés, 2013; X. Wu et al., 2011) had extremely convergent microRNA target sets (for example (G.-Q. Chen et al., 2010) had a predicted target set smaller than expected with p=0). Such exceptionally small target sets mean that any GO term or arbitrary collection of genes hit by chance will contain a significantly high proportion of the target set, causing a large number of significant GO terms to be returned by these 5 input lists.

There have been other proposals to try to harness the convergence of microRNAs and to improve enrichment analysis. The miRSystem tool gives a p-value based on the tendency for the standard method to consistently find the same significant GO terms (T.-P. Lu et al., 2012). By comparing the order of enrichment for a new sample with the order for random microRNAs, outstanding changes can be identified. Although use of a precomputed distribution limits the flexibility of miRSystem, the approach escapes the problem of bias reported here (T.-P. Lu et al., 2012). More recently, methods correcting the problems we report here have been published. A corrected version of the standard method is followed by the tool miEAA (Backes et al., 2016). This webserver uses precomputed annotations of microRNAs, where the label of a GO term is applied whenever one of the targets of that microRNA is associated with the GO term. Input lists of microRNAs for analysis are then analysed by standard enrichment analysis (Backes et al., 2016). The disadvantage of such an approach is that information is lost on how many of the targets of a microRNA or set of microRNAs are associated with the GO term. Therefore, although this method is recommended for analysis, in many

cases it fails to produce useful results, as we find in other sections of this thesis. An alternative is also provided by the project to annotate microRNAs under the Gene Ontology Consortium (Huntley et al., 2016). This will provide curated information on associations with GO terms, which can then be used for further analysis. Improved annotation and better understanding of microRNA functions will allow more reliable interpretation of microRNA differential expression.

4.6. Conclusion

We have highlighted critical problems with the most common general approach to functional enrichment analysis of microRNA target genes. We have shown that testing with the hypergeometric distribution sampling from all GO annotated genes in the genome is inappropriate. Our literature survey showed that a large number of studies reported significant results that are unlikely to stand after correction for the bias in the distribution of targets of randomly sampled microRNAs. We believe that our results provide a strong argument against continued use of the standard method.

5. Building bridges between the autism and research communities: Opinions from people with a lived experience of autism

5.1. Abstract

Autism is a neurodevelopmental condition with a complex spectrum of features, which is still inadequately understood. In recent years, there has been increasing concern that a disconnect between the autism and research communities is harming trust, leading to reduced involvement of autistic people in research, skewed research priorities, queries about research design and data interpretation and a failure to impact the everyday lives of the majority of autistic people. In this work, we present results from a series of workshops held jointly between members of the autism community and researchers aimed at building bridges towards working together better. We report findings from workshop focus group sessions and interviews held with 30 individuals from the autism community. Thematic network analysis of these responses showed key concerns about relations of power and also provided valuable insights into motivations for taking part, and practical recommendations for how to achieve active participation in research. Autistic participants argued strongly that research goals should be pursued together as equals, with autistic people allowed to shape ideas. Participants highlighted missed opportunities to share research findings and to explain the value of projects through a range of media. Participants also requested better information, both initally and throughout, and that researchers consider access requirements more carefully before commencing studies. Altogether, our focus groups suggested real potential for meaningful participation, and showed a strongly positive view that research could have great value to the autism community. We have encapsulated the practical recommendations from our study into a set of guidelines to be made available, which we believe will help autism researchers to build stronger partnerships for successful research in future.

5.2. Background

Autism is a neurodevelopmental condition in which a person has difficulties with social communication and interaction, together with unusually restricted or repetitive behaviour and interests (American Psychiatric Association, 2013). At present, it is believed that approximately 1% of the population is autistic (Y. S. Kim et al., 2011), with a broad spectrum of different experiences and features to each of these people's conditions (Lai et al., 2013).

Academic research into autism has had a fractious and complex relationship with the autistic community in the 92 years since child psychologist Grunya Efimovna Sukhareva first wrote about a condition she called “schizoid psychopathy” in Moscow in 1925, which we would now recognise as 'autism'. In recent years, there has been growing concern about a disconnect between the autism community (autistic adults, parents and family members of autistic people) and the research community. It has been shown that there is a large gap between research priorities identified by academics and funding bodies and those identified by people on the autism spectrum and their families (Pellicano et al., 2014). Concerns about the failure of research to have significant impact on the everyday lives of most autistic people, difficulties with recruitment, and continued distrust within the community all require better engagement to resolve.

There is also a wider debate within academia about participatory research. As discussed by Elsabbagh et. al (Elsabbagh et al., 2014) the demand for research knowledge to have real life impact is conceptualised under different models ranging from a top-down unilateral model in which researchers inform potential users of knowledge (such as practitioners or policy makers) to a much more bi-directional model that involves on-going dialogue between researchers and communities, at all levels of the research cycle. Indeed, although we draw a dichotomy between the autism and

research communities for the purpose of this manuscript, there are in practice many autistic individuals with skills and interests spanning the two communities. This form of sustained and bi-directional engagement is particularly important when considering autism research. Moreover, while development of more ‘participatory’, ‘emancipatory’ or ‘inclusive’ relationships between researchers and communities have been the subject of much debate in the field of learning disability/intellectual disability research (Elsabbagh et al., 2014; McClimens, 2008; Nind & Vinha, 2012), it is still relatively rare within the autism research literature (Chown et al., 2017; Jivraj et al., 2014). Two examples of emerging partnerships in the UK include “Shaping Autism research in the UK” and “Autism: Top 10 Research Priorities” by the charity Autistica.

In this work, we present the findings from a series of workshops which we held together with members of the autism community in the Manchester area, UK. These were conducted as a collaboration between Autism@Manchester (http://www.autism.manchester.ac.uk/), an interdisciplinary network including academics from several research fields, and Salfordautism, an autism support organisation that is led and run by autistic people. The aim of these workshops was to improve mutual understanding and explore how the academic and autism communities could work together better. In this paper, we will use our workshop focus group discussions, questionnaire responses and other collected data to explore the motivation of members of the autism community to engage in research, the power relations between the autism and research communities, and the possibilities and barriers for constructing knowledge together. We end with some practical suggestions about conducting research with the autism community. We have encapsulated the practical recommendations from our study into a set of guidelines to be released online which we believe will help autism researchers to build stronger partnerships for successful research in future.

5.3. Methods

5.3.1. Participants

24 adults (18 years and over) comprising 10 adults with ASC and 14 parents of children with autism took park in the workshops. When applying to take part in the workshop, participants confirmed that they or their child had received a diagnosis of an autism spectrum condition. 17 people took part in the focus groups: 8 participants were autistic adults, 8 participants were parents of autistic children and one was a parent and autistic adult. All participants were based within the Greater Manchester area and were recruited through advertisements in the local community, Disability Support Services at Universities in Greater Manchester, local press and autism support groups in the Greater Manchester area.

Additionally, a further 13 autistic adults were recruited at Autscape following a presentation developed from the discussions at the workshops regarding autistic engagement in research. Autscape is a UK based conference for autistic people which is organised by autistic people. These participants took part in two focus groups and two interviews. They did not contribute to the questionnaire data.

All participants gave informed consent and the study was approved by The University of Manchester Research Ethics committee.

5.3.2. Overview of workshops

A series of three workshops was held, entitled ‘Working Together: Building Bridges between the Autistic and Research Communities’. In Workshop 1 “What is research” participants listened to a presentation from an autistic adult about their experience of participating in different types of research and two presentations from academics detailing the different types of

research methods and ways of getting involved (from participants to co-researchers). In workshop 2 “How can we work together” the focus groups were conducted and we also asked the participants for their ideas on suitable topics for a series of short films where they would be filmed talking about aspects of autism. In focus group 3 “Developing research ideas together” the group listened to an academic presentation outlining the research process (from ideas to generating results) and came up with their own ideas for research which they then entered into a brief mock grant proposal form. The group also chose the film topics they most wanted to discuss and indicated whether they would be willing to feature in these films. These films can be viewed at the Autism@Manchester webpage.

5.3.3. Questionnaire data

At the beginning of workshop 1, participants completed open ended questions on whether they had taken part in research previously and whether they had any concerns about taking part in research and in these workshops. They were also asked their opinions about research involvement using Likert style questionnaires (Figure 5.1).

5.3.4. Focus groups

Three focus groups were run in parallel on the second day of the workshops and facilitated by authors of the paper (EG, AG, TB). The focus groups took approximately one hour and were comprised of 1 parent, 5 autistic adults (focus group 1), 4 parents and 1 autistic adult (focus group 2) and 3 parents, 2 adults and 1 parent/autistic adult (Focus group 3)
The focus group schedule included questions which could be grouped into three areas:

(1) Experiences of having taken part in research previously. For instance, would you participate in research again in the future? If so, how would you like to get involved?

(2) Opinions surrounding research and the participant's perception of the practical relevance of research to their lives. For instance, what advantages do you think there are to participating in research for yourself and for other people?
(3) Opinions on the practicalities involved in taking part in research and of how the research experience might be improved. For instance, how might we make it easier for autistic people and their families to take part in research?

The participants were given access to the schedule in advance of the focus group and were informed that they were welcome to complete written responses to focus group questions. Participants could also send anonymous text messages to an online poll that could be viewed throughout the workshop. This data was also included in the analysis. Participants who were recruited at Autscape completed identical focus group schedules, or interviews, following the same schedule that was used in the workshops.

All focus groups and interviews were recorded and transcribed in full, with participants given a number to preserve their anonymity. Qualitative data from the transcriptions of the focus groups and interviews, as well as written responses was analysed using a thematic network analysis (Attride-Stirling, 2001). Two members of the research team independently coded the data. This was followed by a meeting to compare codes and arrange them under themes. The written analysis was then reviewed by 2 participants in the focus groups.

5.4. Results

5.4.1. Questionnaire responses

50% of autistic adults and parents had participated in research before, covering areas such as neuroscience, sensory issues, anxiety, the use of dogs for autism, genetic, speech and language and education. 12.5% of autistic adult and 0% of parents had concerns about taking part in research. Concerns listed included the worry that data could be misinterpreted or would not be shared. 50% of autistic adults and 33% of parents had concerns about taking part in the workshops. Concerns listed included travel, personal agendas of attendees, stress/anxiety, feeling overwhelmed, emotional impact, childcare and whether they were the correct audience.

Results from the Likert scale questionnaires (Figure 5.1) revealed that participants indicated that they believed autism research had potential relevance to their life and that there was a benefit in the autism and research communities working together. However, the participants had mixed opinions on whether researchers valued their opinions and their knowledge of how to get involved in research and read the findings of research, were mixed.

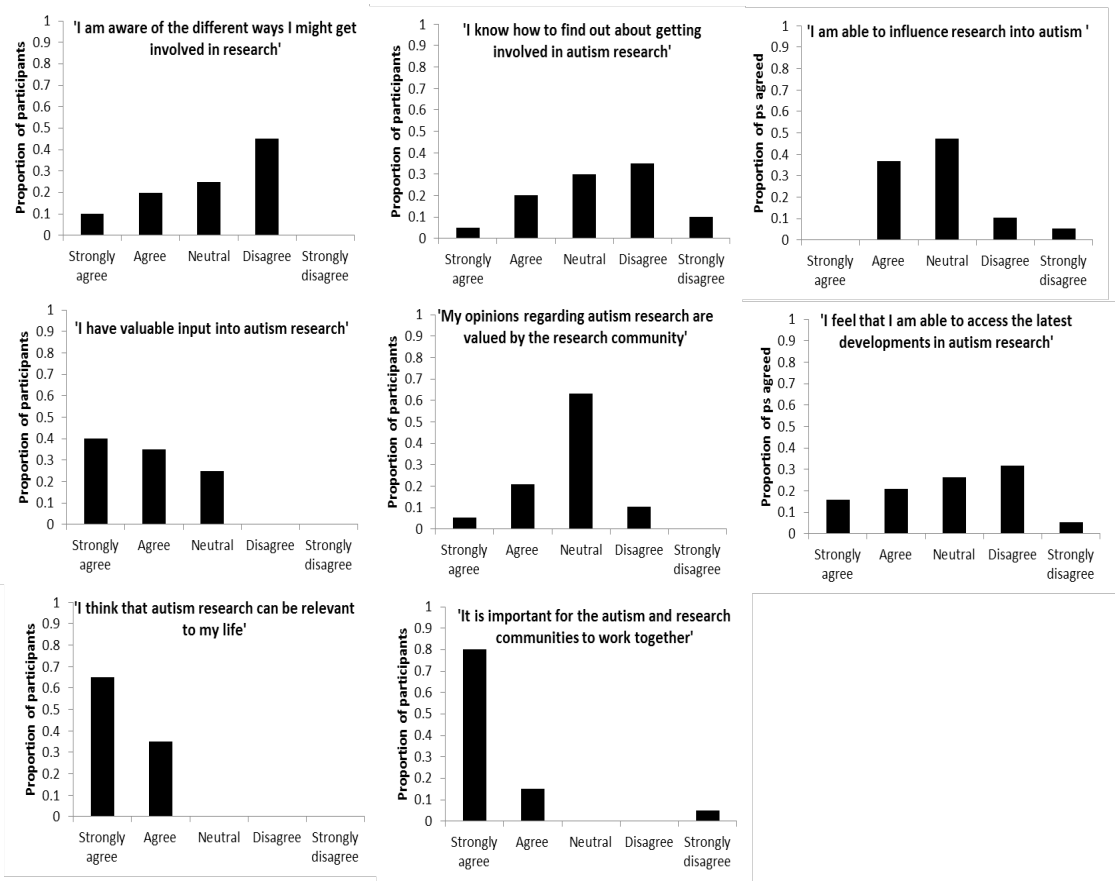


Figure 5.1. Likert scale questionnaire responses

Proportion of participant responses to each item on the questionnaire distributed at the beginning of the workshops (n=20).

5.4.2. Focus group data

Three broad themes were identified in analysis of the focus group discussions – motivation to take part in research, relations of power between the autistim and research communities, and practical advice for designing research projects.

5.4.2.1. Theme 1: motivation to take part in research

Overall the participants in our research were very keen to invest and contribute to research and build stronger ties with the research community. They saw research as potentially beneficial to the autism community and to them as individuals, and all those who had experience of taking part in research have agreed that they will be willing to do this again in the future. However, it should be noted that our participants were necessarily a self-selecting group with willingness to dedicate time to joining the workshops, who were able to communicate in focus groups or take part in interviews, and who thought building relationships with the research community was worthwhile. Behind a broad enthusiasm for taking part in studies, our focus groups told us several key motivations. Among the reasons for engaging with researchers were: a desire for personal learning, altruistic motivations, post-diagnosis support and community, and a desire to direct research towards key targets. Other drivers of participation included free trips to the city of the research institution, and being paid money.

5.4.2.1.1.Personal learning

Participants described enjoying the process of being asked questions at visits, and through explaining things to researchers better understanding themselves.

“One of the advantages is I imagine, would be finding out more about myself.”
Autistic participant, focus group

In particular, learning more about the complexity of what it is to be autistic was highlighted. Similarly, another autistic participant liked finding out her IQ score and was reassured to repeat the ADOS (Lord et al., 2000) and get the same results again. However, researchers do not always do all they can to support this - for example when one focus group member took part in a study he asked for his own personalised outcome information, but was only given overall results.

“I asked them will we get the research back in any detail, because I was really interested to see how I had performed compared to the other people, but they said they'd let everyone know the overall results and conclusions but not the actual individual data for yourself.”
Autistic participant, focus group

While understanding data protection issues, they said they felt that researchers could offer more without jeopardising anonymity for other participants. Research studies give participants other opportunities for learning. One focus group participant told us that she took part in a study about helper dogs, and found the course and training she was given as part of it very helpful. Simply the chance to talk to people and find out what researchers were working on, was also a motivation.

“...just conversations around things relating to autism, I find that really interesting, because I don't like to always talk about autism all the time to most of my neurotypical friends, and I don't want to sound like I'm obsessed with it, but I find it really interesting, so it's nice to have those conversations.”
Autistic participant, focus group

As participation in research was seen as a useful way to learn more about

the self and about mitigating some of the negative effects of life on the autistic spectrum, participants stressed the value of participating in research after receiving their diagnosis. Several commenters talked about poor post-diagnosis support at our focus groups. Post-assessment counseling was slow, detailed information about an individual's condition was not given, and they felt they were “just left to it” according to one focus group participant. Several members of our focus groups commented on the missed opportunity to involve newly diagnosed people in community support networks and research projects. A participant said that they were surprised that after diagnosis they were left without other support.

“I felt happy but at the end of it, you're just handed a pack and that was it.” Autistic participant, focus group

Taking part in research was seen by participants as a way to counter this post diagnosis isolation and confusion. An autistic participant recommended that more information on taking part in research should be given to newly diagnosed people.

5.4.2.1.2. Altruistic motivations

Research participants give up their time, experience stress and share personal information, often with the goal of helping others who might benefit from the research. In general discussion, our focus groups felt that autism was poorly understood by society, that practical research to help autistic people was lacking, and that professionals were often misinformed. Improving this with a wider understanding in society was a primary motivator for some, including in our written interview.

An autistic focus group participant said that she liked to feel useful and to make a difference to someone else. Other participants more specifically spoke of being motivated because autism is serious for society and should be understood better.

“I'm sure you all feel it's something that is a serious thing that ought to be understood better and studied more and diagnosed more accurately in its various forms and ramifications, and that we should contribute to work with people who are doing that, to make sure the process of understanding goes on well. That's something we can and ought to do.”
Autistic participant, focus group

Focus groups also talked about wanting to make a contribution, to move forwards and give something back. Our workshop members who were newly diagnosed were particularly motivated with a 'social responsibility'.

Participants spoke of a desire to take part in research for the benefit of their children. One parent in our focus group wanted to learn about the underlying genetic condition for her daughters.

“So in my view, the more knowledge I have, the more I'm likely to be able to help my daughter, so it's for those reasons really.”
Parent of autistic child, focus group

Another focus group participant similarly told us he hoped that some ideas might help his son.

5.4.2.1.3. Shaping research goals

Our participants agreed that autistic people’s involvement in research must go far beyond assuming the passive role of research participants and had a strong desire to influence research targets and express their wishes for research into neglected areas. Previous studies have shown a marked difference in research priorities of the autism community with actual present research funding (Pellicano et al., 2014).

Some of our female participants told us that they felt particularly

misunderstood and dissatisfied with the lack of research into female autism. One autistic member told us that she would like more research on females on the spectrum, commenting that she had been seen by psychologists for 16 years before an autism diagnosis. One of our participants who has an autistic daughter and Asperger's herself was so motivated that she organised and ran a study on Asperger's in females. A female parent of an autistic child raised the problem that research she had seen was not very relevant to what she was going through.

"...what I feel as a parent as well is that any research I had come across, it's not very relevant to what we're going through or what we want to see... I asked my friends to see if there's anything that they wanted to see, and one of them was female diagnosis and another one was about twins."
Parent of autistic child, focus group

A focus on basic research may have missed working on more practical help which could make a big difference to the day-to-day life of parents. These comments suggest that some of the weakness in engagement between researchers and the autism community could be greatly improved if research goals were more directly inspiring to the community. Moreover, the demand that research goals should be much more accountable to the autism community's identified needs and agendas was not only seen as a way to make research more applicable and beneficial to the lives of autistic people and their families, but crucially, also as a way to ward off some of what they saw as more harmful and even sinister implications of research. Many of the participants indicated that while they were keen to take part in research they were concerned that some of the research being carried out could actually be negative for autistic people. The main concern was around research that promotes the idea that autism is a form of pathology that needs to be eradicated or cured.

"...if I learn later that the researcher is actually pro- cure or that sort of thing then I would be very uncomfortable about having participated."
Autistic participant, Autscape session

This objection to cure fits the discourse of the neurodiversity movement, a movement that has emerged in the last 20 years through the activism of adults on the autistic spectrum. The three main principles in this discourse of autism is that 1) autism is the result of a cognitive neurological (rather than psychological) difference (Ortega, 2009), 2) that this difference is of value and should not be seen as inferior, tragic or pathological (Chamak, 2008) and 3) that the hardships associated with autism are more often the result of social barriers, negative attitudes and mutual misunderstandings, rather than an inherent result of the condition (Milton, 2012). Further, this understanding of autism meant that participants recognised the value of bio-medical and laboratory research aimed at identifying the biological, neurological and cognitive mechanisms that underlie autism, which were seen as potentially contributing to increased self-awareness (see earlier section on personal learning) and to increasing professional and public understandings of autism. However, while this form of research was seen as valuable, participants also warned that this research must be accompanied by much more social and cultural consideration, to avoid a pathologising and cure seeking approach.

"I would think there may be a difficulty because you are only focusing on the medical side and it might be better to look at the social side and the cultural side and the identity, cultural identity as well. And solutions aren't, probably aren't going to be found through drugs."
Autistic participant, Autscape session

This stress on the need for social research had, for our participants, two main reasons:

1. Participants stressed that many of the problems and issues faced by autistic people and their family members happen in social contexts, and could well be the result of social expectations or administrative demands, and that such issues may be impossible to identify under controlled conditions. Thus, social research is not only a way to

identify such issues but also a way to improve implementation of research knowledge into services and policy

"I think it was very clear in the discussion last week at the end that people are mainly concerned about all those structural and practical issues that happen in their lives, so research should be able to address that and have a realistic plan for how to not just study it, but actually impact"
Autistic participant and parent of autistic child, focus group

2. Some forms of cultural and social research that pay attention to the specific ideas of autistic people were seen as a way of recognising autistic identities as well as enabling and supporting autistic people to participate in valued social roles and activities. However, as participants stressed, this function of supporting and enabling positive autistic identities and participation is not inherent to social research as such. Indeed, such research, if not conducted with respect for the autistic individual's own self-understanding can be a form of "cultural imposition". Thus, having autistic perspectives, and indeed autistic people, involved in all levels of autism research, including the design, interpretation and dissemination, was stressed by our participants as central to building bridges between the autism and research communities and to conducting research that is accountable to the needs, priorities and aspirations of the autism community. It is these issues of collaboration and dialogue that we will now turn to.

5.4.2.2. Theme 2: equalising power relations and constructing knowledge together

One of the key issues that were raised by participants was the lack of involvement of the autistic community in shaping and disseminating autism research. It is worth pointing out again that overall participants were positive about their experiences of taking part in research, praising care

taken with confidentiality, and being treated with respect and sensitivity. However, the unequal power relations of research production were seen as a major obstacle for collaboration and a source of mistrust and dissatisfaction with autism research.

Participants often felt that their involvement in research was limited to a passive role of being a source of data, and being asked to perform tasks without much explanation or feedback.

"...research is mainly conducted for the researchers' sake, who are pursuing their research questions, with not enough attention given to giving back."
Autistic participant and parent of autistic child, focus group

"I think the downfall from my point of view, or any participant's point of view, is that you are not able to get feedback. Otherwise you […] thank you, here's your money- bugger off."
Autistic participant, Autscape session

5.4.2.2.1. Sharing research findings

An area of great concern for workshop participants was the availability (or rather lack thereof) of research knowledge to the autism community. Participants criticised a failing to share outcomes quickly, with scientists being 'precious' about their findings. This included first and foremost autistic people, particularly those newly diagnosed (see section on motivation). However, participants have also stressed the need to have better dissemination of research findings to groups such as practitioners and policy makers.

"*I just feel really simple things, if we're aiming to build bridges, how can you possibly find out about those sorts of materials because they're only available, all that information is only available to you […] There is very much that sense that we don't have access to any of this information or any*

of the resources or decisions. That makes me feel very excluded at times and I think that everyone else is saying similar things"
Autistic participant, focus group

This exclusion from research findings and knowledge was attributed by participants to three main factors: difficulty in locating relevant research, difficulty in accessing research papers due to pay-wall restrictions, and difficulties in understanding research papers written in specialised jargon with genre conventions. Targeted write-ups for a lay audience or for professionals, alongside standard academic papers, might be a way to improve publishing practices in the area.

"I mean you could write something of 2 or 3 pages that could go to autism services, for example, that came out of this, that doesn't need to be published. So there's lots of other things you can do, and that's taking responsibility for filtering the information you consider relevant to particular service providers and disseminating it in a form that is easy to understand and will make sense to that service provider."
Autistic participant, focus group

Participants wanted to see research findings shared in accessible forms and a variety of medias. Many participants discussed the need to present research findings through short videos and animations which could be shared and publicised on websites and social media sites. These were seen as more accessible than research papers, and suitable to a variety of audience including those who struggle with reading. However, as one participant stressed, it is important to include closed captions on such videos for the benefit not only of those with hearing impairment, but also of the many people on the autistic spectrum with auditory processing difficulties.

Social media was stressed as a useful platform for sharing research findings (whether written or film) also because of its wide reach and the ability to comment and engage in discussions. The significance of the internet to the emergence and growth of the neurodiversity movement and autistic

community has been stressed by many writers, both from within and outside the movement (Ortega, 2009; Robertson & Ne'eman, 2008), as a space that provides opportunities for autistic people to come together over distance, and without facing the (often overwhelming demands) of crowded gatherings and the need to communicate with strangers in real time. Yet, as our participants pointed out, not everybody is on the internet for a variety of reasons and it is important to also hold dissemination events and seminars in physical space.

5.4.2.2.2. Overcoming barriers to sharing information

Participants showed an awareness of some of the institutional and paradigmatic barriers that may lead to allocating participants a passive role. For instance, limited time and money available to researchers (required to process results and inform participants) as well as certain methodological designs that require research participants to remain ignorant of the real topic of the research. Yet, they felt that even in such cases participants should be informed in advance that there is a hidden element to the research and be informed of the true aim, and their own results, as soon as possible. This was seen as an important element of showing honesty and building trust, without which participants may feel deceived or even, in the case of one participant, suspect that some clandestine research was conducted on them without their knowledge. As a minimum, participants expected to be fully informed about the aims of research and their individual results.

5.4.2.2.3. Meaningful active participation in research

However, participants in our workshops went well beyond this minimum in their envisaging of the desired relationships between the autism and research communities, to demand much greater participation of autistic

people in the shaping and dissemination of research. The ability to *discuss* research findings was seen as crucial for building bridges between the autism and research communities. First, participants have stressed that a more elaborate and nuanced understanding of the meaning and implication of research findings can be gained from discussion as opposed to just reading a report or listening to a lecture. Second, they valued opportunities to discuss research findings with researchers also as a way of making researchers more aware of autistic ideas and interpretations of the findings. In other words, workshop participants did not see dissemination as the final stage of research in which the researcher delivers their findings to the uninformed. Instead they offered a more reciprocal and cyclical view, in which the researchers share their findings and receive feedback that may alter the interpretations of results or spring new ideas for research, a process similar to the idea of peer reviews and discussions within the research community. Thus, similar to the demand to include autistic people in designing research and setting priorities, this view of dissemination challenges the existing power relations with academic knowledge production and asks that researchers will be more accountable to the autism community. Participants suggested that in future, researchers should endeavour to build up community credibility through their engagement. Similar understanding of community engagement in neurodiversity research was called for by Elsabbagh et al. (Elsabbagh et al., 2014) who outline the changes in models of knowledge exchange from a focus of unilateral flow of information from scientists to potential users of this knowledge, toward active communication with and engagement of knowledge users and researchers.

The active and meaningful participation of autistic people in shaping research aims and questions was also seen as crucial for ensuring that autism research is aligned to the needs and aspirations of the autism community.

“I think that humility is tremendously important. Recognising your own comparative ignorance when you embark on a research study is hugely

important, recognising that at that stage you might be well advised to keep your research questions as open ended as possible. Your participants may tell you some things which cause your study to go off in directions you may never have anticipated." Autistic participant, Autscape individual interview

"*But you know, if they, rather than sort of traditionally, historically, have things done to them, like being examined under the microscope, if they can actually say this is the sort of thing we think needs researching, and actually have more input and more output from the process as well."*
Parent of autistic child, focus group

While researchers hold sophisticated methodological knowledge needed for designing valid data collection and interpretation tools, autistic people hold unique expertise in day to day life with autism, as well as in the cultures and discourses of the autism community (Milton, 2012), and must therefore play a significant role in devising research questions and agendas that have positive impact on the lives of people on the autism spectrum. This was aptly put by one of the participants:

"From my own point of view, I do feel that research should not simply be undertaken from the perspective of experts who have some kind of vaulted higher position from which they can examine the rest of the world. And I feel that that's an incredibly important reason why I'm participating."
Autistic participant, focus group

It is worth noting here that while in this paper we refer to the autism community and the research community as separate groups, several of the participants in our workshop had formal experience in conducting research across a range of disciplines (all but one not directly linked to autism research), ranging from writing dissertations, to conducting PhD research and holding paid research positions. This overlap suggests that involving (at least a certain part of) the autism community in all stages of the research may be less difficult than sometimes imagined. Further, autistic people with expertise in both research and autism culture and community may be

significant players in building bridges between academics and autistic people with no formal research training, support the development of more inclusive methods for data collection and dissemination, and help build skills and empower autistic people to take a greater part in research.

5.4.2.3. Theme 3: practical advice for involving the autistic community in research

Through the Building Bridges project, we worked with participants to develop practical advice for how to involve the autistic community in research projects. We have divided these suggestions into four key areas: recruitment, information, access and running sessions.

5.4.2.3.1. Recruitment

Recruitment of participants for research projects can be a major difficulty for academics. Apart from delayed or cancelled projects, recruitment difficulties can also undermine the validity of research. For example, small sample sizes in neuroscience have meant that average statistical power is low, and that it is less likely for statistically significant results to reflect true effects (Button et al., 2013). Limits to recruitment also bias what studies can be performed, and for autism research may explain why some areas are understudied.

Participants in our focus groups generally recognised that recruitment to studies was far from optimal. Several expressed a desire to take part more, and regretted that they did not know how to.

“I would be happy, but again I think it comes back to have you ever seen any advertising for it.”
Parent of autistic child, focus group

Although clearly our participants were self-selected for willingness and ability to participate in a study, there was broad enthusiasm amongst those who took part for more participation and greater recruitment to studies. However, suggestions for how to improve recruitment were less uniform.

Participants discussed use of social media, its benefits and difficulties, and potential for recruitment to studies. Some participants wanted more presence from researchers on Facebook, for example suggesting that results of autism research should be reported on Facebook groups, and that Facebook's event invite system could be used to invite participants to studies.

One participant who herself performed a research study was very successful in finding online survey participants through social media.

“I think that's the beauty of social media. It's easier to contact... one site in particular has over 20,000 members and several others have a few hundred. So I was able to quite quickly get enough responses.”
Parent of autistic child, focus group

Other participants spoke of preferring email communication to phone, because of less stress and having the chance to think. However, several participants also reflected on problems with social media. A basic problem for researcher engagement is that many people are not on Facebook, and any scheme of sharing results or recruiting for studies through social media is likely to only target a specific community. This risks excluding those unwilling or unable to join those groups. For example, one participant spoke of 'missing out'.

Another problem with social media and online groups is the possibility of emotionally charged and aggressive discussions. A focus group participant specifically discussed preferring support groups offline because of such encounters. Similarly, a parent of an autistic child said she would prefer 'face-to-face' communication, and another participant also talked about

having to step in to stop heated discussions as an admin for an online group. Scientists and researchers may also be put off from better engagement, both offline and online, because of fears about hostility.

Support groups and societies may solve some of the issues with online engagement and recruitment. Several participants recommended advertisement of opportunities through the National Autistic Society (NAS). Participants discussed researchers having a bigger presence on the NAS website. Advertising on local support groups was also mentioned, with one participant recommending the local group Salfordautism. In fact, much of our recruitment for the Building Bridges workshops was done through our links with Salfordautism. We found that these links made it much easier to maintain long-term and stable connections with the local autism community. This made it possible for example to re-contact the network for further projects such as producing videos and collaborating on this article.

Our written interview participant described a problem of adverts and opportunities being scattered, meaning more effort for researchers and potential participants missing finding what they wanted. He suggested having a single system to look for adverts. A focus group discussion also suggested that autism networks could be better integrated with each other. Specifically, a participant wanted work to develop a directory of support group contacts for the autism community.

“I don't know what research is being done in terms of collating what groups there are for people on the spectrum or for parents with kids on the spectrum, if there's like a comprehensive directory of contacts. Because I think it would be nice to have some kind of network or framework where for example you have the research community, people in the community running their own support groups, you have the NHS and so on.”
Autistic participant, focus group

A directory with opportunities for research participation could help to link different groups together, and reduce work trying to find out about research

opportunities.

Local schools may be a good starting point for bringing together networks and advertising to the local community. Local doctors are also a preferred direct connection suggested in our written interview. One parent of an autistic child was able to take part in a research study without directly visiting the researchers, as she was recruited by her consultant and simply sent a DNA sample. She was frustrated with the slowness, which may have been exacerbated by lacking direct contact, but was sympathetic to the reasons.

A major problem with many of the recruitment methods commonly used, and those suggested at our workshops, was a bias against those who have more difficulty communicating.

"...there are a few people on there who really suffered because they went on it and they were very shy or they were less verbal and they had a much harder time communicating. And I think that their opinions on research and their value to the research process would be invaluable... You're asking us but we're only a small representative group, we're not representative of the whole autistic community by any stretch of the imagination and I would like everyone to be represented. I think that's very important. But also for the research, it'd be very valuable."
Autistic participant, focus group

The focus group also spoke of the value of their input, which is sadly missing, as those who take part cannot truly be representative of the whole diverse community.

Some other examples of people missing out on studies include an autistic participant forgetting to book onto a study they were due to take part in, and a parent having no opportunities at all in India, where he never saw any advert for research. Informed consent is another barrier to participation, where researchers may be dissuaded from engagement by complications to

getting approval for involving children and people with learning disabilities.

5.4.2.3.2. Information

Autistic people tend to have a preference for clear plans and can find change stressful (Lai et al., 2013). Indeed, participants put less emphasis on information about study aims and background in comparison with practical information to improve access and reduce anxiety.

Researchers sometimes fail to provide adequate advance information to participants. Our focus groups described the stress caused by experiences of having to change appointment times, and the need for better detailed information from their participation. Our focus group participants were clear that this was an important issue. Two autistic participants both said that information about what would be involved in the study was important to feel confident beforehand.

“I think sometimes it's also not knowing what the research involves, and it could be that it involves the child but the parent doesn't know hat's actually involved, what's needed. So maybe if some information was given that we're going to discuss this. And also a chance for the parent and the child, or the person to come with somebody and have a look around, so they know what they're going into.”
Autistic participant, focus group

This included details that some researchers might not anticipate, such as photos of the room that the study would take place in. A participant also praised a local pantomime that offered the chance to go and visit the room in advance as well as providing an information pack with pictures.

Failing to properly provide information may have surprising consequences for researchers as well as causing stress and discouraging participation. One participant described taking part in a study which failed to mention that they

would do an autism assessment. This bothered him, and so he gave a 'psychopathic story' in answer to one of the questions. Information must be given beforehand, and it should be comprehensive. It must include details about the schedule for the visit, and change or additions to the schedule should be avoided. Detailed photos and travel instructions should also be sent. Another participant praised good information provided for a study that she took part in, and it was clear that this helped to give her a good experience.

"I thought it was really good actually. We got information before we went about the study and what to expect when we got there. The hardest thing was actually finding the room, it always seems to be an issue! ...But the actual study was well set up and we got told some information about it, and they were lovely with the children and very welcoming."
Parent of autistic child, focus group

Participants also complimented our information packs for the Building Bridges workshops specifically for including pictures. A participant said that the reason pictures are useful is that they give a good idea of what situation is likely to be encountered. This can then be used to predict likely problems in advance, such as a noisy foyer.

5.4.2.3.3. Access

Obstacles to participants who have been recruited to join a study include immediate issues with getting into and navigating through buildings, as well as more abstract issues such as life commitments and money. Despite a positive experience and good prior information, a participant reported difficulties in finding and getting to the room for the study she took part in. She told us that this was often an issue, and made harder by bringing children. The building was also not easy to access with a pram. Her access to the study could have been improved by having more help getting in. Clearly researchers should check for pram access, and also ensure

wheelchair access. This was also raised by an autistic participant who was frustrated by directions that failed to take into account disability.

“I never managed to get to any of the tutorials because nobody could describe how to get to the tutorial in such a manner that a disabled person could actually get there.”
Autistic participant, focus group

She felt that the people giving directions had not properly checked, by walking through themselves, so that the directions could be followed. In general, we were told that travelling and getting to studies was stressful by several participants. One focus group member knew of university researchers travelling out themselves to meet participants, which could potentially solve these access problems, although also raising other issues.

For our workshops, we tried to ease stress by putting several signs up in the building, as well as the photos of the venue sent beforehand. Participants said that these helped them. We also had researchers waiting at the building entrance to welcome participants and show them on to our room, and who stayed for latecomers. One participant said this was 'fantastic'. In addition, one member waited at the local train station and then walked from there to the venue with participants who requested this help. An autistic participant said that without this, she would not have been able to make her way to the venue.

“That made a massive difference because if you'd just have given me a map, I would not have found my way here and I would have panicked.”
Autistic participant, focus group

Generally these were not particularly onerous preparations for the research team to make for these workshops, and the benefits in reducing stress and widening participation are clear. It should therefore be common practice where possible to hold research in buildings that can be accessed more easily with prams and wheelchairs, with clear signage, and members of the

research team meeting participants where necessary.

For participants physically able to reach venues for research, personal constraints on time and cost may instead prevent access. For one autistic participant, these were the most important considerations for taking part. Again, it was highlighted that the actual research question was not as important as practical issues with making time to go.

"I think an important thing to note is that my considerations for this were nothing to do with the actual research, they were more to do with my availability to come and be researched upon. So it was time constraints, diarising, it was obviously the cost, etc, and I'm balancing all of those things."
Autistic participant, focus group

This participant did not have children, but clearly childcare issues could exacerbate this problem. For one parent, arranging childcare and time to come to a study was the biggest reason to stop her from taking part. For people with older children, the time at which a study is held can be difficult, as one mother needed to finish in time to collect her daughter after school. Holding the Building Bridges workshops, we attempted to find a suitable time, balancing busy traffic on the streets, and choosing a time that parents could make. We chose to begin our sessions at noon. However, there were still issues, as students leaving lectures at lunchtime were noisy, and there was pressure at the end of the day for people with commitments. Some focus group members suggested Saturday or other weekend appointments as a potentially better solution.

5.4.2.3.4. Running sessions

Autistic participants of research studies have needs that are diverse and which researchers should try to meet where possible. Clearly part of the challenge for researchers wanting to widen participation is to provide

positive experiences that will encourage the community to engage more. Poor practice may also lead to unreliable data.

For example, an autistic participant described how questionnaires with loose questions dependent on time could cause confusion.

“And when I first got there I was feeling quite stressed and wound up and agitated and I answered the questionnaire accordingly. But half an hour later when I'd done the first experiment, I felt much more relaxed and I told the researchers that if you gave me that questionnaire again, my answers would be very different.”
Autistic participant, focus group

His answers would have changed completely if given a little time to relax, and may have caused contradictions with answers given later in the day. At the least, for answers not reflective of recent stress travelling to a venue, it is sensible to have some time to recover before starting a study. One autistic participant suggests 'a cup of tea and a quick chat and a bit of chill out time before you start doing an experiment' to help with the anxiety. This participant also raised the issue of confusion with questionnaires and tests. She told us that she had taken the ADOS (Lord et al., 2000) multiple times in studies, and has ended up trying to keep her answers the same as in her first test, because she recognises that she has evolved over time, and her answers could otherwise also depend on mood. Researchers must be aware that confusion about tests like this may bias results, and must attempt to be very explicit about what they want with each question they ask.

Tiredness of participants may similarly lead to changes in response through a long session. An autistic participant described finding studies tiring sometimes. Part jokingly she told us, 'There better be biscuits.' Generally our participants asked that researchers be sympathetic to difficulties, and provide breaks and drinks where needed.

For the running of focus groups, participants emphasised the need for clear

moderation with objective boundaries given before group discussions and ways to move on the conversation from one speaker. Participants recognised the difficulty with having many people wanting to talk freely. Against this, researchers must balance the desire to not excessively hold discussion on their own terms preventing participants from expressing themselves.

5.4.3. Discussion

In this work, we have reported the findings of a series of workshops aimed at building bridges between the autism and research communities. Following thematic network analysis, we identified three themes discussed by focus group participants. On motivations and reasons to take part in research, participants highlighted chances for personal learning, post-diagnosis support, altruistic motivations and a desire to shape research goals. Discussing power relations, participants felt that involvement of the autism community in research was too passive, that research findings were not made available, and argued that barriers to meaningful participation should be broken down. We identified practical advice for improving involvement in studies, including better recruitment approaches, access considerations and recommendations for running sessions.

5.4.3.1. Trust and sharing

Trust and sharing between the autism and research communities were issues which spanned all of the themes that we covered. Participants revealed major distrust about the direction of past and present scientific research, as other studies of the disability movement have found (Chamak, 2008). They emphasised the need to be treated as equals in the pursuit of research, including being allowed to help shape ideas. There was a frustration that academics were not willing to listen and really hear disagreements with ways of seeing things, as highlighted previously among an Asperger's consultation group (Chown et al., 2017). The participants felt that researchers had, through a long record of weak engagement, failed to

explain their work or help their research impact lives of autistic people (Elsabbagh et al., 2014). Participants were keen to promote a wider understanding of autism among professionals working in autism, the research community and the wider population. In particular, they wanted to challenge assumptions from academics, highlighting that the findings and assumptions of researchers hold weight and can have real world impact on the experiences of autistic people. However, through clearer and more accessible information about studies together with better sharing of research findings and opportunities to get involved, our focus groups proposed positive steps to improve trust and participation.

5.4.3.2. Engagement as an ongoing process

Overall, there was a strong positive view about research with participants understanding and supporting its value. However, our focus group analysis showed that engagement was necessary throughout the research process. Participants highlighted the importance of including the autistic perspective throughout the research process. From study design to implementation, dissemination and discussions moving forward, the autism community suggested opportunities for participation. Our study also highlighted the value of long-term engagement with support groups to achieve such ongoing participation. Our experience of working with Salfordautism for these workshops gave us the chance to build deeper relationships over time, establishing trust and mutual understanding. Our participants also felt that involvement in research provides a level of support such as the opportunity to talk about autism and to find out about the latest developments in studies. In line with this they indicated a need to better link post-diagnosis support groups with research involvement opportunities and findings.

Altogether, our focus groups suggested great potential for meaningful participation, and showed a strongly positive view that research could have value to the autism community. We believe that building bridges like this will help to mobilise for change in wider institutional cultures and in the

funding of research. In 2014, a survey of literature identified only seven studies describing such participatory research in autism or other neurodevelopmental conditions (Jivraj et al., 2014). However, since then other groups have undertaken participatory projects, including “Shaping Autism research in the UK” which aimed to present seminars with the perspectives of researchers, practitioners and the autism community on equal terms. Ongoing inclusive research in learning disabilities is similarly developing interdependency, negotiation and support (Nind & Vinha, 2012). The findings from this study have galvanised a series of initiatives at Autism@Manchester including the development of an Expert by Experience group and a regular newsletter. In addition, we have encapsulated the practical recommendations from our study into a set of guidelines which will help autism researchers to build stronger partnerships for successful research in future.

5.4.3.3. Ongoing challenges

This work has highlighted potential for building bridges between the autism and research communities in a way which might facilitate participatory research. Nevertheless, there remain ongoing challenges that limit this work and pose a barrier to participatory research in autism (McClimens, 2008). The imbalance in access to materials, such as through academic libraries, and knowledge and skills means that it is not always straightforward for academics to communicate findings or justify practice. The imbalance in power between communities also means that autistic individuals have difficulties despite desiring greater involvement. We attempted to reduce bias in this work caused by researchers undertaking the thematic network analysis by including a review stage involving two autistic participants who raised issues which may otherwise have been undiscussed (Attride-Stirling, 2001). From the research community perspective, an additional barrier is that the incentive structure in many research disciplines does not directly include participatory research. In fact, the time taken may be at the expense of work more beneficial to an academic career. To encourage more

researchers to adopt better practice including the autism community in research, institutions could find ways of rewarding this work.

Different ways in which autistic and non-autistic people express themselves can create barriers to participation (McClimens, 2008). We found that dialogue at workshops could break down where the discussion strayed from the topic, or where behaviour or comments made progress difficult. Participants noted that the researchers set the agenda and thus the discussion will tend to be based around the topics which are important to the researchers. Furthermore, participants raised frustration in the autism community that communication deficits are always seen as lying with the autistic person. This in turn tends to shape the focus of much research. Our project was necessarily limited in scope as all participants were verbal and capable of responding to a call for participants. Therefore, although we refer to the views of 'the autism community' for simplicity here, it should be noted that this is biased towards only a subset of all autistic individuals, parents and family members. It will be important to explore avenues which can enable all autistic voices to be heard in the research process in future.

5.4.3.4. Reflection on implications for other studies in this thesis

Many of the points raised by participants of the Building Bridges workshops are closely related to other work which the author has undertaken in this thesis. In particular, members raised concerns about eugenics and doubts about the motivation of genetics research. Through taking part in the workshops, I heard from autistic individuals who specifically mentioned mistrust of Autism Speaks and the MSSNG Project. Reflecting on these discussions, I attempted to present motivations for my work more clearly and restrict the scope of conclusions. In the analysis of microRNAs as potential biomarkers, these discussions led me to more careful assessment of accuracy based on utility for early diagnosis and intervention.

Participants had mixed views on the value of more basic research. Although many recognised the importance of such work, it would certainly have been helpful to have engaged earlier. Such engagement may have raised ideas and views that could have led to valuable additions to the project. Better engagement with participants may have also helped to give a more grounded understanding of individual phenotypes. The experience of the workshops showed that binary labelling of samples by autism status masked a large amount of difference between individuals relevant to their conditions.

6. Conclusions

6.1. Thesis overview

In the work presented in this thesis, we have addressed the role of microRNAs in autism and undertaken a critical assessment of research methods in this field. We highlighted six key targets for investigation in our introduction, covering (1) rare variants in microRNAs, (2) differential expression, (3) post-transcriptional regulation, (4) microRNAs as biomarkers for autism, (5) computational approaches to microRNA analysis and (6) community engagement. These targets were centered on the goal of improving research practice for autism studies and microRNA computational analysis, as well as advancing knowledge on microRNA regulation in autism. We performed four studies to investigate these key areas. Using small RNA sequencing for 42 autism LCL samples and 10 control samples, we tested for differential expression of microRNAs, potential biomarkers and post-transcriptional regulation of microRNAs. We then expanded a search for rare variants in microRNAs to whole genome sequencing data for 671 probands from the MSSNG Project (Yuen et al., 2015). We investigated bias in microRNA functional enrichment analysis and flaws in a commonly applied method. We organised a series of Building Bridges Workshops with 30 focus group and interview participants to establish best practice for engaging with the autism community and producing and disseminating autism research more effectively. Our core aims led to the following advancements and findings. (1) Although we did not find significant rare variants in seed regions in the small RNA sequencing dataset, our expanded search in MSSNG data identified 101 such SNVs. (2) Our study found differentially expressed microRNAs, and these provided both confirmation of some previous results as well as novel candidates with strong evidence for regulatory pathways in autism. (3) We developed methods for the analysis of post-transcriptional regulation of microRNAs, identifying differential arm usage, 5' isomiRs and A-I editing in autism. (4) We obtained reasonable accuracy for autism status predictions

based on microRNA expression, but the limited dataset meant that we could not advance our work for diagnostic use. (5) We found bias that led to a large number of scientific reports of pathways enriched for targets of microRNAs in diseases and biological states which in practice were not targeted more than expected by chance. (6) We identified problems in ongoing autism research due to poor dissemination and weak engagement with the autism community, and proposed practical solutions to address these issues.

6.2. Differential expression of microRNAs in autism

Our investigation into differential expression of microRNAs in our autism samples revealed 24 significantly differentially expressed microRNAs. We also identified 8 microRNAs with significant arm switching, and 24 5' isomiRs with differential expression in autism. There was strong overlap between our results and previous reports for relevant microRNAs. These suggested promising modes of microRNA regulation for investigation in autism.

We found several microRNAs which had previously been reported as involved in conditions related to autism phenotypes, including schizophrenia (hsa-miR-132-5p, hsa-mir-146a and hsa-miR-181b-5p), neurofibromatosis type 1 (hsa-let-7b-5p) and recovery after stroke (hsa-miR-181b-5p) (Beveridge et al., 2008; Chan & Kocerha, 2012; Masliah-Planchon et al., 2013; Z. Peng et al., 2013; Zou et al., 2012). How apparently similar genetic causes can alternatively lead to intellectual disability, autism, schizophrenia, or other neurodevelopmental disorders, remains one of the key unresolved questions in autism research (S. H. Lee et al., 2013). Clarifying similarities and differences at a transcriptome level may therefore help distinguish developmental pathways for these different disorders.

We proposed possible modes of control of the microRNA profile in autism based on previous reports on the microRNAs which were found affected in our study. These included DNA methylation changes (hsa-miR-142-3p), regulated differential expression across different brain regions (hsa-miR-191-5p), and complex post-transcriptional regulation of arm usage and alternative 5' isomiRs (hsa-miR-181b-5p and hsa-mir-146a) (Mor et al., 2015; Ziats & Rennert, 2013). Better basic understanding of how cell conditions regulate microRNA pathways, such as through DNA methylation and target abundance effects (Kang, Choi, Hong, & Lee, 2013), are necessary for a clearer picture of microRNA control in autism.

We found microRNAs affected in our study which suggest possible mechanisms through which dysregulation could contribute to autism development. We identified microRNAs known to be associated with neural inflammation (hsa-miR-23a/b-3p) and dendritic growth changes (hsa-mir-146a and hsa-miR-132-5p) (Jin et al., 2004; Nguyen et al., 2016; Zhu et al., 2012). We further found possible explanations for these effects through previous reports of associations with APT1 (hsa-mir-138) and FMRP (hsa-miR-132-5p, hsa-miR-125b-5p, hsa-let-7c-5p and hsa-mir-138) (Edbauer et al., 2010; Siegel et al., 2009). Connections to FMRP provide a particularly strong mechanism for dysregulation in autism, as the etiology of Fragile X syndrome is already well established. Further confirmation here of multiple dysregulated microRNAs with strong FMRP interactions shows that this pathway is important more broadly in non-syndromic autism, and that treatment approaches developed for Fragile X syndrome may have potential for broader application (Bagni et al., 2012).

6.3. Discovery of post-transcriptional regulation

In order to find differential isomiR and arm usage, we used an algorithm which first counted reads aligned at each alternative 5' start site for mature microRNAs. We then performed a Mann-Whitney U test for significant differences in ratios between arm read counts and between dominant and

other 5' isomiR read counts. Using this test may give less sensitivity than modelling the distribution of read counts for 5' isomiRs and then performing a statistical test using this model (Anders & Huber, 2010), as information is lost calculating ratios. However, the conservative approach still revealed significant arm switching for key microRNAs hsa-mir-181b and hsa-mir-146a. Arm switching has previously been reported in different insect species, among different tissues, and at different developmental stages (S. Li et al., 2012; Marco et al., 2010). Our test here was able to detect smaller changes in arm usage, as we detected a shift in ratios of usage even where there was no switch in which arm was dominant. Arm selection preference changes have also been reported in cancer for hsa-mir-193a (Tsai, Leung, Lo, Chen, & Chan, 2016). It is not fully understood how Ago and other involved proteins select the arm for incorporation into RISC, although thermodynamic features of the microRNA duplex affect the usage (Meijer, Smith, & Bushell, 2014). A mechanism for dynamic regulation based on cell mRNA expression has been proposed where high target gene abundance may protect passenger arms from degradation (Chatterjee, Fasler, & Bu, 2011). This was consistent with observations in breast cancer arm usage (Tsai et al., 2016). Similar study of luciferase activity for RNA constructs with artificial binding sites for microRNAs identified in our study with differential arm usage may clarify whether such a mechanism may drive these effects in autism (Tsai et al., 2016).

We tested for differential proportions of A-I editing and for differential 5' isomiR usage in autism, but did not find significant changes. However, it has been reported previously that dominant 5' isomiR usage can change through evolution, across cell types and in different developmental stages (Fernandez-valverde, Taft, & Mattick, 2010; Tan et al., 2014). Further studies using methods trialed here may provide robust statistical tests to reveal 5' isomiR regulation in other biological states. Previous reports have also found differential A-I editing in mRNAs in cancer and psychiatric disorders (Slotkin & Nishikura, 2013). It has been shown that ADAR1 dysregulation in cancer leads to dysregulation of microRNA involved in cancer pathways, but differential A-I editing for microRNA has not yet been

demonstrated (Nemlich et al., 2013). Therefore, quantification of reads at greater depth targeted at sites with known A-I editing and testing for changes in different biological states may provide new avenues for understanding microRNA post-transcriptional regulation. Sequencing projects with greater read depth may also make discovery of new A-I editing sites possible.

6.4. MicroRNA variants

The nature of autism heritability is unresolved, with studies showing evidence both for common SNPs having major contributions (Klei et al., 2012), as well as evidence for rare and *de novo* variants driving autism (Luo et al., 2012). We investigated the hypothesis that multiplex autism cases may follow the model proposed where rare variants are inherited leading to convergent disruption on key pathways (Zhao et al., 2007). Through Sanger sequencing, we confirmed that there were no rare seed region variants in our original dataset. An expanded search in the MSSNG whole genome sequencing dataset identified 101 such variants in 671 probands, but none of these were *de novo* variants. Prediction of target properties of the variant mature microRNAs suggested that they may affect radically different gene sets, supporting a hypothesis of microRNAs causing broad dysregulation. However, weaknesses in target prediction software (Ritchie et al., 2009), and a failure to find enriched functional classes means that the consequences of these SNVs and their relevance to autism are difficult to quantify.

6.5. Potential of microRNAs as biomarkers for autism

One of the aims of our project was to explore potential for using microRNA expression to predict autism status. We used LCL expression as a proxy for expression in the brain, based on previous work showing a good correspondence in results for the two (Talebizadeh et al., 2008).

Development of such a model in older autistic individuals also requires consistent signatures for toddlers for useful early intervention (Walsh et al., 2011). Expression profiles of immune related genes have previously been reported as modestly predictive (75% accuracy) for autism status (Pramparo et al., 2015). The expression of microRNAs in saliva was also used previously to obtain 95% prediction accuracy (Hicks et al., 2016), however these results are unlikely to be robust on unseen test data, as the microRNAs used to build the predictive model were selected based on differential expression in all samples, including the test set. In our study, we achieved comparable accuracy to those previous reports (81%), but poor performance for control samples. It is clear that the weakness in our approach was a lack of control samples for model training. However, the work demonstrates that microRNAs can in principle be used to predict autism status. Further studies using larger training sets may build towards reliable early prediction. This may be improved with the addition of other sample information such as polygenic risk score and sex (Robinson et al., 2016). Better modelling of genetic predictors of autism status may also help the understanding of how autism etiology relates to phenotypes in the general population (Clarke et al., 2016).

6.6. Critical assessment of functional enrichment analysis methods

While developing methods for understanding the functional implications of microRNA dysregulation in our work, we found critical problems with a large number of published studies in this field. Functional enrichment analysis is often used with datasets of targets of differentially expressed microRNAs. The popular standard method of analysis typically returned hundreds of enriched categories, from which authors were free to select appealing items while neglecting enrichments that did not provide an easy explanation. We believe that as well as the immediate methodological bias that we identified in our report, this raises wider issues in scientific practice. In particular, a focus on producing publications based on significant p-

values may lead researchers to report findings that are less likely to be true (Ioannidis, 2005). This issue was also raised during our Building Bridges project focus groups, where systemic pressures in academic research led to distrust in the autism community. Since we reported on the bias in standard microRNA functional enrichment analysis (Bleazard et al., 2015), alternative methods have been developed which perform enrichment analysis at the level of microRNAs to avoid bias (Backes et al., 2016). However, other studies continue to use biased analyses, including reports on microRNAs in autism (Hicks et al., 2016).

6.7. Recommendations for participatory research

Autism genetics, development and biomarkers are controversial research topics. Successful investigation of these areas requires collaborative approaches and engagement with the autism community (Walsh et al., 2011). Because of this, we explored how best to involve the autism community in research through the Building Bridges workshops. Thematic network analysis of focus groups showed strong support for scientific efforts in general, including basic research such as the work reported here, but also raised several relevant concerns. Motivations and goals for autistic participants in research included shaping research towards practical goals (Pellicano et al., 2014). Discussion of power relations with the scientific community led to recommendations that autistic individuals and carers be involved in study design from the start. In practice, in our research into microRNAs in autism, we did not engage with the community to the extent that we have recommended here. However, results from the Building Bridges workshop have informed our analysis in setting more practical parameters for interpretation of results, and in improving dissemination of findings. This supports our critique of bias in functional enrichment analysis, attempting to reduce false discoveries that reduce trust, as well as more stringent criteria for machine learning success in our work. The writing of the chapter reporting our Building Bridges work also reflects the collaborative approach trialled here. Based on dialogue with autistic

participants, we attempted to write including their perspective and highlighting issues they wanted us to raise. However, the breadth of opinions in the autism community that we were able to convey was limited by motivation and ability to participate, and likely missed the voice of parents and individuals with more severe impairments. We believe that our work will lead to improved practice more broadly in autism research, through our guidelines into participant access, information, recruitment, dissemination, active participation and practical requirements.

6.8. Concluding remarks

Together, we hope the work presented in this thesis will improve methods for microRNA analysis and autism research, and that discoveries of microRNA differential expression and post-transcriptional regulation lead to better understanding of affected networks in autism.

Supplementary materials

Supplementary Table 2.1. List of top candidates for 5' isomiR usage change

Mature product region information	Mann-Whitney U-test 2-sided p-value
('hsa-miR-30d-5p', 'MIMAT0000245', '-', 135817183, 'chr8')	0.0003593638
('hsa-miR-150-5p', 'MIMAT0000451', '-', 50004110, 'chr19')	0.0005103334
('hsa-miR-342-5p', 'MIMAT0004694', '+', 100576010, 'chr14')	0.0013263888
('hsa-miR-4256', 'MIMAT0016877', '-', 113004445, 'chr1')	0.0013949533
('hsa-miR-24-3p', 'MIMAT0000080_1', '+', 97848346, 'chr9')	0.0022510147
('hsa-miR-27a-3p', 'MIMAT0000084', '-', 13947281, 'chr19')	0.0038994229
('hsa-miR-146a-5p', 'MIMAT0000449', '+', 159912379, 'chr5')	0.0045814577
('hsa-miR-484', 'MIMAT0002174', '+', 15737158, 'chr16')	0.006275276
('hsa-miR-103a-3p', 'MIMAT0000101', '+', 3898188, 'chr20')	0.0068273939
('hsa-miR-103a-3p', 'MIMAT0000101_1', '-', 167987931, 'chr5')	0.0068273939
('hsa-miR-301b', 'MIMAT0004958', '+', 22007314, 'chr22')	0.0072090894
('hsa-miR-16-5p', 'MIMAT0000069', '-', 50623184, 'chr13')	0.0074728249
('hsa-miR-363-3p', 'MIMAT0000707', '-', 133303433, 'chrX')	0.0087845703
('hsa-miR-155-5p', 'MIMAT0000646', '+', 26946295, 'chr21')	0.0108054983
('hsa-miR-624-3p', 'MIMAT0004807', '-', 31483895, 'chr14')	0.0109700946
('hsa-miR-1273c', 'MIMAT0015017', '+', 155174503, 'chr6')	0.0112052867
('hsa-miR-192-5p', 'MIMAT0000222', '-', 64658695, 'chr11')	0.0123168341
('hsa-miR-664a-3p', 'MIMAT0005949', '-', 220373913, 'chr1')	0.0126367026
('hsa-miR-197-3p', 'MIMAT0000227', '+', 110141562, 'chr1')	0.0136992917
('hsa-miR-222-3p', 'MIMAT0000279', '-',	0.0161192081

Mature product region information	Mann-Whitney U-test 2-sided p-value
45606462, 'chrX')	
('hsa-miR-6511a-3p', 'MIMAT0025479', '+', 15019837, 'chr16')	0.0161998481
('hsa-miR-6511a-3p', 'MIMAT0025479_3', '-', 18437893, 'chr16')	0.0161998481
('hsa-miR-6511a-3p', 'MIMAT0025479_2', '+', 16462776, 'chr16')	0.0161998481
('hsa-miR-6511a-3p', 'MIMAT0025479_1', '+', 16418488, 'chr16')	0.0161998481
('hsa-miR-331-5p', 'MIMAT0004700', '+', 95702221, 'chr12')	0.017661614
('hsa-miR-466', 'MIMAT0015002', '-', 31203228, 'chr3')	0.0176803792
('hsa-miR-532-5p', 'MIMAT0002888', '+', 49767773, 'chrX')	0.0180159424
('hsa-miR-186-3p', 'MIMAT0004612', '-', 71533346, 'chr1')	0.0195640552
('hsa-miR-339-3p', 'MIMAT0004702', '-', 1062613, 'chr7')	0.0220432835
('hsa-miR-425-5p', 'MIMAT0003393', '-', 49057654, 'chr3')	0.0235851764

Supplementary Table 4.1. List of functional enrichment analysis studies

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
Characterization of global microRNA expression reveals oncogenic potential of miR-145 in metastatic colorectal cancer.	Yes	Human	MAMI	1
Differential expression of microRNAs in adipose tissue after long-term high-fat diet-induced obesity in mice.	No	Mouse	miRWalk	10

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
Systematic analysis of microRNA involved in resistance of the MCF-7 human breast cancer cell to doxorubicin.	Yes	Human	MAMI	176
Co-modulated behavior and effects of differentially expressed miRNA in colorectal cancer.	Yes	Human	MetaCore known targets	15
High-throughput sequencing reveals differential expression of miRNAs in intestine from sea cucumber during aestivation.	Yes	Sea Cucumber	TargetScan	9
miR-124 regulates adult neurogenesis in the subventricular zone stem cell niche.	No	Mouse	TargetScan	1
Microvesicles derived from adult human bone marrow and tissue specific mesenchymal stem cells shuttle selected pattern of miRNAs.	Yes	Human	miRWalk	11
MicroRNA-218 is deleted and downregulated in lung squamous cell carcinoma.	No	Human	PicTar; TargetScan; miRBase; miRNAMap	1
Potentially important microRNA	No	Human	miRGen	2

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
cluster on chromosome 17p13.1 in primary peritoneal carcinoma.				
Identification and characterization of microRNAs by high through-put sequencing in mesenchymal stem cells and bone tissue from mice of age-related osteoporosis.	Yes	Mouse	Mireap	8
Detection of microRNA Expression in Human Peripheral Blood Microvesicles	Yes	Human	miRanda; TargetScan	1
Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival.	Yes	Human	miRgate	18
Abnormal microRNA expression in Ts65Dn hippocampus and whole blood: contributions to Down syndrome phenotypes.	Yes	Mouse	microRNA.org; mirDB; RNA22; TargetScan; microT	12
UVA and UVB Irradiation Differentially Regulate microRNA Expression in Human Primary Keratinocytes	Yes	Human	miRBase; PicTar; TargetScan	9
Deep Sequencing of Small RNA	Yes	Mouse	TargetScan; miRTarBase; miRecords;	37

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
Repertoires in Mice Reveals Metabolic Disorders-Associated Hepatic miRNAs			TarBase	
Brain and blood microRNA expression profiling of ischemic stroke, intracerebral hemorrhage, and kainate seizures	No	Rat	MAMI	5
Computational analysis of microRNA function in heart development	Yes	Mouse	miRanda; TargetScan; PITA; MirTarget2; RNAhybrid; TarBase; miRecords	27
Intrinsic Features in MicroRNA Transcriptomes Link Porcine Visceral Rather than Subcutaneous Adipose Tissues to Metabolic Risk	Yes	Pig	PicTar; TargetScan; MicroCosm Targets	22
Sequencing and bioinformatics-based analyses of the microRNA transcriptome in hepatitis B-related hepatocellular carcinoma.	No	Human	miRBase; TargetScan	3
Transcriptome profiling of microRNA by Next-Gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk	No	Human	TargetScan; miRanda	10

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
Direct regulation of microRNA biogenesis and expression by estrogen receptor beta in hormone-responsive breast cancer	No	Human	TargetScan	67
MicroRNA profiling of peripheral nerve sheath tumours identifies miR-29c as a tumour suppressor gene involved in tumour progression	No	Human	PITA; TargetScan	5
MicroRNA classifiers for predicting prognosis of squamous cell lung cancer	No	Human	MAMI	15
Identification and pathway analysis of microRNAs with no previous involvement in breast cancer	Yes	Human	miRanda; TargetScan; PicTar	130
De-regulated microRNAs in pediatric cancer stem cells target pathways involved in cell proliferation, cell cycle and development	Yes	Human	BcmicrO	26
Neuronal MicroRNA Deregulation in Response toAlzheimer's Disease Amyloid-beta	Yes	Mouse	TargetScan	1
MicroRNA-30c reduces hyperlipidemia and atherosclerosis in mice by	No	Human	TargetScan	1

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
decreasing lipid synthesis and lipoprotein secretion				
Unraveling the global microRNAome responses to ionizing radiation in human embryonic stem cells	Yes	Human	miRanda	39
Expression Patterns of MicroRNAs in Porcine Endometrium and Their Potential Roles in Embryo Implantation and Placentation	Yes	Pig	miRanda; TargetScan; miRDB	21
MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours	No	Human	TarBase; miRTarBase; miRecords; Ingenuity Systems Knowledge database	46
MicroRNA Transcriptome in Swine Small Intestine during Weaning Stress	Yes	Pig	TargetScan	98
miRNA expression profile analysis in kidney of different porcine breeds.	No	Pig	DIANA-microT	8
Genome-wide analysis of microRNA and mRNA expression signatures in hydroxycamptothecin-resistant gastric cancer cells.	No	Human	miRGen	25
Pseudorabies virus infected porcine	Yes	Pig	miRanda	20

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs.				
Identification and Differential Expression of microRNAs in Ovaries of Laying and Broody Geese (Anser cygnoides) by Solexa Sequencing.	Yes	Goose	Unclear	367
Cellular microRNAs 498 and 320d regulate herpes simplex virus 1 induction of Kaposi's sarcoma-associated herpesvirus lytic replication by targeting RTA.	Yes	Human	RNAHybrid; PITA; FindTar	109
Identification of hsa-miR-335 as a prognostic signature in gastric cancer.	Yes	Human	TargetScan	1
MicroRNA dysregulation in the spinal cord following traumatic injury.	Yes	Rat	MiRNA-Map	53
MicroRNA expression profiling in HCV-infected human hepatoma cells identifies potential anti-viral targets induced by interferon-α.	Yes	Human	Jtarget	1
Identification of differentially expressed	Yes	Human	TargetScan; miRDB	25

Title	Mentioned in Abstract	Species	Prediction Algorithms	MicroRNA List Extracted
microRNAs across the developing human brain.				

References

Abelson, J. F., Kwan, K. Y., O'Roak, B. J., Baek, D. Y., Stillman, A. a, Morgan, T. M., … State, M. W. (2005). Sequence variants in SLITRK1 are associated with Tourette's syndrome. *Science*, *310*(5746), 317–20. doi:10.1126/science.1116502

Abrouk, M., Zhang, R., Murat, F., Li, A., Pont, C., Mao, L., & Salse, J. (2012). Grass microRNA gene paleohistory unveils new insights into gene dosage balance in subgenome partitioning after whole-genome duplication. *The Plant Cell*, *24*(5), 1776–92. doi:10.1105/tpc.112.095752

Abu-Elneel, K., Liu, T., Gazzaniga, F. S., Nishimura, Y., Wall, D. P., Geschwind, D. H., … Kosik, K. S. (2008). Heterogeneous dysregulation of microRNAs across the autism spectrum. *Neurogenetics*, *9*(3), 153–61. doi:10.1007/s10048-008-0133-5

Alon, S., Mor, E., Vigneault, F., Church, G. M., Locatelli, F., Galeano, F., … Eisenberg, E. (2012). Systematic identification of edited microRNAs in the human brain. *Genome Research*, *22*(8), 1533–40. doi:10.1101/gr.131573.111

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)* (p. 1000). American Psychiatric Publishing.

Ander, B. P., Barger, N., Stamova, B., Sharp, F. R., & Schumann, C. M. (2015). Atypical miRNA expression in temporal cortex associated with dysregulation of immune, cell cycle, and other pathways in autism spectrum disorders. *Molecular Autism*, *6*(37), 1–13. doi:10.1186/s13229-015-0029-9

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. doi:10.1186/gb-2010-11-10-r106

Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., … Bourgeron, T. (2012). Individual common variants exert weak effects on the risk for autism spectrum disorders. *Human Molecular Genetics*, *21*(21), 4781–92. doi:10.1093/hmg/dds301

Applied Biosystems. (2013). Overview of SOLiD^{TM} Sequencing Chemistry. Retrieved from http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html

ASDWGPGC. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism*, *8*(21), 1–17. doi:10.1186/s13229-017-0137-9

Ashburner, M., Ball, C., Blake, J., & Botstein, D. (2000). Gene Ontology:

tool for the unification of biology. *Nature Genetics*, *25*(May), 25–29.

Attride-Stirling, J. (2001). Thematic networks: an analytic tool for qualitative research. *Qualitative Research*, *1*(3), 385–405.

Axtell, M. J., Westholm, J. O., & Lai, E. C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology*, *12*(221), 1–13.

Backes, C., Khaleeq, Q. T., Meese, E., & Keller, A. (2016). miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Research*, *44*(W1), W110–W116. doi:10.1093/nar/gkw345

Bagatell, N. (2010). From Cure to Community : Transforming Notions of Autism Historical Trends Influencing the Emergence of an Autism Community. *Ethos: Journal of the Society for Psychological Anthropology*, *38*(1), 33–55. doi:10.1111/j.1548-1352.2009.01080.x.34

Bagni, C., Tassone, F., Neri, G., & Hagerman, R. (2012). Science in medicine Fragile X syndrome: causes, diagnosis, mechanisms, and therapeutics. *Journal of Clinical Investigation*, *122*(12). doi:10.1172/JCI63141.4314

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: evidence from a British twin study. *Psychological Medicine*, *25*(1), 63–77.

Baker, J. P. (2013). Autism at 70 -- redrawing the boundaries. *The New England Journal of Medicine*, *369*(12), 1089–91. doi:10.1056/NEJMp1306380

Barbash, S., Shifman, S., & Soreq, H. (2014). Global Coevolution of Human MicroRNAs and Their Target Genes. *Molecular Biology and Evolution*, *31*(5), 1237–1247. doi:10.1093/molbev/msu090

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, *136*(2), 215–33. doi:10.1016/j.cell.2009.01.002

Basu, S. N., Kollu, R., & Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Research*, *37*(Database issue), D832–6. doi:10.1093/nar/gkn835

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., … Koller, D. (2013). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, *24*, 14–24. doi:10.1101/gr.155192.113

Bedford, R., Jones, E. J. H., Johnson, M. H., Pickles, A., Charman, T., & Gliga, T. (2016). Sex differences in the association between infant markers and later autistic traits. *Molecular Autism*, *7*, 21. doi:10.1186/s13229-016-0081-0

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300.

Berg, J., & Geschwind, D. (2012). Autism genetics: searching for specificity

and convergence. *Genome Biology*, *13*(7), 247.

Beveridge, N. J., Tooney, P. a, Carroll, A. P., Gardiner, E., Bowden, N., Scott, R. J., … Cairns, M. J. (2008). Dysregulation of miRNA 181b in the temporal cortex in schizophrenia. *Human Molecular Genetics*, *17*(8), 1156–68. doi:10.1093/hmg/ddn005

Bizuayehu, T. T., Lanes, C. F. C., Furmanek, T., Karlsen, B. O., Fernandes, J. M. O., Johansen, S. D., & Babiak, I. (2012). Differential expression patterns of conserved miRNAs and isomiRs during Atlantic halibut development. *BMC Genomics*, *13*(11).

Bleazard, T., Lamb, J. a, & Griffiths-Jones, S. (2015). Bias in microRNA functional enrichment analysis. *Bioinformatics (Oxford, England)*, *31*(10), 1592–8. doi:10.1093/bioinformatics/btv023

Bonev, B., Stanley, P., & Papalopulu, N. (2012). MicroRNA-9 Modulates Hes1 Ultradian Oscillations by Forming a Double-Negative Feedback Loop. *Cell Reports*, *2*(1), 10–18. doi:10.1016/j.celrep.2012.05.017

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., … Shipley, G. L. (2009). The MIQE Guidelines : Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, *622*, 611–622. doi:10.1373/clinchem.2008.112797

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*. doi:10.1038/nrn3475

Buxbaum, J. D., Daly, M. J., Devlin, B., Lehner, T., Roeder, K., & State, M. W. (2012). The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron*, *76*(6), 1052–6. doi:10.1016/j.neuron.2012.12.008

Buyske, S. (2009). Comment on the article “Heterogeneous dysregulation of microRNAs across the autism spectrum” by Abu-Elneel et al. *Neurogenetics*, *10*(2), 167. doi:10.1007/s10048-009-0179-z

Carter, M. T., & Scherer, S. W. (2013). Autism spectrum disorder in the genetics clinic: a review. *Clinical Genetics*, *83*(5), 399–407. doi:10.1111/cge.12101

Caudy, A. A., Myers, M., Hannon, G. J., & Hammond, S. M. (2002). Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes & Development*, *16*(516), 2491–2496. doi:10.1101/gad.1025202.GENES

Chamak, B. (2008). Autism and social movements: French parents’ associations and international autistic individuals' organisations. *Sociology of Health & Illness*, *30*(1), 76–96. doi:10.1111/j.1467-9566.2007.01053.x

Chan, A. W. S., & Kocerha, J. (2012). The Path to microRNA Therapeutics in Psychiatric and Neurodegenerative Disorders. *Frontiers in Genetics*, *3*, 82. doi:10.3389/fgene.2012.00082

Chatterjee, S., Fasler, M., & Bu, I. (2011). Target-Mediated Protection of Endogenous MicroRNAs in C. elegans. *Developmental Cell*, *7*, 388–396. doi:10.1016/j.devcel.2011.02.008

Cheloufi, S., Dos Santos, C. O., Chong, M. M. W., & Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*, *465*(7298), 584–9. doi:10.1038/nature09092

Chen, G.-Q., Zhao, Z.-W., Zhou, H.-Y., Liu, Y.-J., & Yang, H.-J. (2010). Systematic analysis of microRNA involved in resistance of the MCF-7 human breast cancer cell to doxorubicin. *Medical Oncology*, *27*, 406–15. doi:10.1007/s12032-009-9225-9

Chen, M., Zhang, X., Liu, J., & Storey, K. B. (2013). High-throughput sequencing reveals differential expression of miRNAs in intestine from sea cucumber during aestivation. *PloS One*, *8*(10), e76120. doi:10.1371/journal.pone.0076120

Chepelev, I., Wei, G., Tang, Q., & Zhao, K. (2009). Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Research*, *37*(16), e106. doi:10.1093/nar/gkp507

Cho, S., Jang, I., Jun, Y., Yoon, S., Ko, M., Kwon, Y., … Lee, S. (2013). MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic Acids Research*, *41*(Database issue), D252–7. doi:10.1093/nar/gks1168

Chown, N., Robinson, J., Beardon, L., Downing, J., Hughes, L., Leatherland, J., … Macgregor, D. (2017). Improving research about us, with us: a draft framework for inclusive autism research. *Disability & Society*, *32*(5), 720–734. doi:10.1080/09687599.2017.1320273

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, *11*(6), 415–425. doi:10.1038/nrg2779

Clarke, T.-K., Lupton, M. K., Fernandez-Pujals, A. M., Starr, J., Davies, G., Cox, S., … McIntosh, A. (2016). Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Molecular Psychiatry*, *21*(3), 419–425. doi:10.1038/mp.2015.12

Cloonan, N., Wani, S., Xu, Q., Gu, J., Lea, K., Heater, S., … Grimmond, S. M. (2011). MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biology*, *12*(126).

Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., … Bolton, P. (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA Psychiatry*, 1–9. doi:10.1001/jamapsychiatry.2014.3028

Consortium, T. O. T. G. P. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. doi:10.1038/nature15393

Crawley, J. N. (2012). Translational animal models of autism and neurodevelopmental disorders. *Dialogues in Clinical Neuroscience*,

14(3), 293–305.

Dai, X., & Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research*, *39*(Web Server issue), W155–9. doi:10.1093/nar/gkr319

Davalos, V., & Esteller, M. (2010). MicroRNAs and cancer epigenetics: a macrorevolution. *Current Opinion in Oncology*, *22*(1), 35–45. doi:10.1097/CCO.0b013e328333dcbb

Davidson, M. R., Larsen, J. E., Yang, I. a, Hayward, N. K., Clarke, B. E., Duhig, E. E., … Fong, K. M. (2010). MicroRNA-218 is deleted and downregulated in lung squamous cell carcinoma. *PloS One*, *5*(9), e12560. doi:10.1371/journal.pone.0012560

De Oliveira, L. F. V., Christoff, A. P., & Margis, R. (2013). isomiRID: a framework to identify microRNA isoforms. *Bioinformatics*, *29*(20), 2521–2523. doi:10.1093/bioinformatics/btt424

Devlin, B., Melhem, N., & Roeder, K. (2010). Do common variants play a role in risk for autism? Evidence and theoretical musings. *Brain Research*, *1380*, 78–84. doi:10.1016/j.brainres.2010.11.026

Doench, J. G., & Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes & Development*, 504–511. doi:10.1101/gad.1184404.species

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, *38*(Web Server issue), W64–70. doi:10.1093/nar/gkq310

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, Doyle, F., … Kaul, R. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Dweep, H., Sticht, C., Pandey, P., & Gretz, N. (2011). miRWalk – Database : Prediction of possible miRNA binding sites by "'walking'" the genes of three genomes. *Journal of Biomedical Informatics*, *44*(5), 839–847. doi:10.1016/j.jbi.2011.05.002

Edbauer, D., Neilson, J. R., Foster, K. a, Wang, C.-F., Seeburg, D. P., Batterton, M. N., … Sheng, M. (2010). Regulation of synaptic structure and function by FMRP-associated microRNAs miR-125b and miR-132. *Neuron*, *65*(3), 373–84. doi:10.1016/j.neuron.2010.01.005

Elsabbagh, M., Yusuf, A., Prasanna, S., Ruff, C. A., & Fehlings, M. G. (2014). Community engagement and knowledge translation: Progress and challenge in autism research. *Autism*, *18*(7), 771–781. doi:10.1177/1362361314546561

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biology*, *5*(1), R1. doi:10.1186/gb-2003-5-1-r1

Evans, B. (2013). How autism became autism: The radical transformation of a central concept of child development in Britain. *History of the Human Sciences*, *26*(3), 3–31. doi:10.1177/0952695113484320

Fabian, M. R., Cieplak, M. K., Frank, F., Morita, M., Green, J., Srikumar, T., … Sonenberg, N. (2011). miRNA-mediated deadenylation is orchestrated by GW182 through two conserved motifs that interact with CCR4–NOT. *Nature Structural & Molecular Biology*, *18*(11). doi:10.1038/nsmb.2149

Fernandez-valverde, S. L., Taft, R. J., & Mattick, J. S. (2010). Dynamic isomiR regulation in Drosophila development. *RNA*, *16*, 1881–1888. doi:10.1261/rna.2379610.RNA-Induced

Finotello, F., & Camillo, B. Di. (2014). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, *14*(2), 130–142. doi:10.1093/bfgp/elu035

Fiore, R., Khudayberdiev, S., Christensen, M., Siegel, G., Flavell, S. W., Kim, T.-K., … Schratt, G. (2009). Mef2-mediated transcription of the miR379-410 cluster regulates activity-dependent dendritogenesis by fine-tuning Pumilio2 protein levels. *The EMBO Journal*, *28*(6), 697–710. doi:10.1038/emboj.2009.10

Friedman, R., & Farh, K. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*, 92–105. doi:10.1101/gr.082701.108.

Garbett, K., Ebert, P. J., Mitchell, A., Lintas, C., Manzi, B., Mirnics, K., & Persico, A. M. (2008). Immune transcriptome alterations in the temporal cortex of subjects with autism. *Neurobiology of Disease*, *30*(3), 303–11. doi:10.1016/j.nbd.2008.01.012

Gardiner, E., Beveridge, N. J., Wu, J. Q., Carr, V., Scott, R. J., Tooney, P. a, & Cairns, M. J. (2012). Imprinted DLK1-DIO3 region of 14q32 defines a schizophrenia-associated miRNA signature in peripheral blood mononuclear cells. *Molecular Psychiatry*, *17*(8), 827–40. doi:10.1038/mp.2011.78

Garg, S., Green, J., Leadbitter, K., Emsley, R., Lehtonen, A., Evans, D. G., & Huson, S. M. (2013). Neurofibromatosis Type 1 and Autism Spectrum Disorder. *Pediatrics*, *132*(6).

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., … Devlin, B. (2014). Most genetic risk for autism resides with common variation. *Nature Publishing Group*, *46*(8), 881–885. doi:10.1038/ng.3039

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., … Thompson, A. (2013). Investigating phenotypic heterogeneity in children with autism spectrum disorder : a factor mixture modeling approach. *The Journal of Child Psychology and Psychiatry*, *2*(54), 206–215. doi:10.1111/j.1469-7610.2012.02588.x

Geschwind, D. H. (2011). Genetics of autism spectrum disorders. *Trends in Cognitive Sciences*, *15*(9), 409–16. doi:10.1016/j.tics.2011.07.003

Ghahramani Seno, M. M., Hu, P., Gwadry, F. G., Pinto, D., Marshall, C. R., Casallo, G., & Scherer, S. W. (2011). Gene and miRNA expression

profiles in autism spectrum disorders. *Brain Research*, *1380*, 85–97. doi:10.1016/j.brainres.2010.09.046

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., van de Vorst, M., van Bon, B. W. M., Willemsen, M. H., … Veltman, J. a. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature*. doi:10.1038/nature13394

Gilman, S. R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M., & Vitkup, D. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, *70*(5), 898–907. doi:10.1016/j.neuron.2011.05.021

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., … Schier, A. F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. *Science*, *308*(5723), 833–8. doi:10.1126/science.1109020

Girardot, M., Cavaillé, J., & Feil, R. (2012). Small regulatory RNAs controlled by genomic imprinting and their contribution to human disease. *Epigenetics*, *7*(12), 1341–8. doi:10.4161/epi.22884

Gotham, K., Marvin, A. R., Taylor, J. L., Warren, Z., Anderson, C. M., Law, P. a, … Lipkin, P. H. (2015). Characterizing the daily life, needs, and priorities of adults with autism spectrum disorder from Interactive Autism Network data. *Autism : the international journal of research and practice*, *19*(7), 794–804. doi:10.1177/1362361315583818

Gupta, S., Ellis, S. E., Ashar, F. N., Moes, A., Bader, J. S., Zhan, J., … Arking, D. E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nature Communications*, *5*, 1–8. doi:10.1038/ncomms6748

Gusev, Y. (2008). Computational methods for analysis of cellular functions and pathways collectively targeted by differentially expressed microRNA. *Methods*, *44*(1), 61–72. doi:10.1016/j.ymeth.2007.10.005

Gusev, Y., Schmittgen, T. D., Lerner, M., Postier, R., & Brackett, D. (2007). Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. *BMC Bioinformatics*, *8*, S16. doi:10.1186/1471-2105-8-S7-S16

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software : An Update. *SIGKDD Explorations*, *11*(1), 10–18.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., … Risch, N. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of General Psychiatry*, *68*(11), 1095–102. doi:10.1001/archgenpsychiatry.2011.76

Hamam, R., Hamam, D., Alsaleh, K. A., Kassem, M., Zaher, W., Alfayez, M., … Alajez, N. M. (2017). Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death and Disease*, *8*(9), 1–12. doi:10.1038/cddis.2017.440

Hansen, T., Olsen, L., Lindow, M., Jakobsen, K. D., Ullum, H., Jonsson, E.,

… Werge, T. (2007). Brain expressed microRNAs implicated in schizophrenia etiology. *PloS One*, *2*(9), e873. doi:10.1371/journal.pone.0000873

Hara, Y., Ago, Y., Takano, E., Hasebe, S., Nakazawa, T., Hashimoto, H., … Takuma, K. (2017). Prenatal exposure to valproic acid increases miR-132 levels in the mouse embryonic brain. *Molecular Autism*, *8*(33), 1–9. doi:10.1186/s13229-017-0149-5

Hervas, A., Matthews, N., Palferman, S., Wallace, S., Aubin, A., Michelotti, J., … Park, D. C. (1998). A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Human Molecular Genetics*, *7*(3), 571–578.

Hicks, S. D., Ignacio, C., Gentile, K., & Middleton, F. A. (2016). Salivary miRNA profiles identify children with autism spectrum disorder, correlate with adaptive behavior, and implicate ASD candidate genes involved in neurodevelopment. *BMC Pediatrics*, *16*(52), 1–11. doi:10.1186/s12887-016-0586-x

Hicks, S. D., & Middleton, F. A. (2016). A Comparative Review of microRNA Expression Patterns in Autism Spectrum Disorder. *Frontiers in Psychiatry*, *7*(November), 1–10. doi:10.3389/fpsyt.2016.00176

Hoen, P. A. C., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., … Lappalainen, T. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nature Biotechnology*, *31*(September), 1015–1022. doi:10.1038/nbt.2702

Hong, D., Rhie, A., Park, S.-S., Lee, J., Ju, Y. S., Kim, S., … Seo, J.-S. (2012). FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics*, *28*(5), 721–3. doi:10.1093/bioinformatics/bts023

Hu, V. W. (2013). From genes to environment: using integrative genomics to build a “systems-level” understanding of autism spectrum disorders. *Child Development*, *84*(1), 89–103. doi:10.1111/j.1467-8624.2012.01759.x

Hu, V. W., & Steinberg, M. E. (2009). Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. *Autism Research*, *2*(2), 67–77. doi:10.1002/aur.72

Hu, Z., Yu, D., Gu, Q.-H., Yang, Y., Tu, K., Zhu, J., & Li, Z. (2014). miR-191 and miR-135 are required for long-lasting spine remodelling associated with synaptic long-term depression. *Nature Communications*, *5*, 3263. doi:10.1038/ncomms4263

Huang, D. W., Sherman, B. T., & Lempicki, R. a. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. doi:10.1038/nprot.2008.211

Huang, F., Long, Z., Chen, Z., Li, J., Hu, Z., Qiu, R., … Jiang, H. (2015). Investigation of Gene Regulatory Networks Associated with Autism

Spectrum Disorder Based on MiRNA Expression in China. *PloS One*. doi:10.1371/journal.pone.0129052

Humphreys, D. T., & Suter, C. M. (2013). miRspring: a compact standalone research tool for analyzing miRNA-seq data. *Nucleic Acids Research*, *41*(15), e147. doi:10.1093/nar/gkt485

Huntley, R. P., Sitnikov, D., Orlic-milacic, M., Balakrishnan, R., Eustachio, P. D., Gillespie, M. E., … Auken, K. V. A. N. (2016). Guidelines for the functional annotation of microRNAs using the Gene Ontology. *RNA*, *22*, 667–676. doi:10.1261/rna.055301.115.

IMGSAC. (2001). A Genomewide Screen for Autism : Strong Evidence for Linkage to Chromosomes 2q , 7q , and 16p. *American Journal of Human Genetics*, (69), 570–581.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8). doi:10.1371/journal.pmed.0020124

Jiang, H., & Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, *25*(8), 1026–32. doi:10.1093/bioinformatics/btp113

Jin, P., Zarnescu, D. C., Ceman, S., Nakamoto, M., Mowrey, J., Jongens, T. a, … Warren, S. T. (2004). Biochemical and genetic interaction between the fragile X mental retardation protein and the microRNA pathway. *Nature Neuroscience*, *7*(2), 113–7. doi:10.1038/nn1174

Jivraj, J., Sacrey, L., Newton, A., Nicholas, D., & Zwaigenbaum, L. (2014). Assessing the influence of researcher–partner involvement on the process and outcomes of participatory research in autism spectrum disorder and neurodevelopmental disorders: A scoping review. *Autism*, *18*(7), 782–793. doi:10.1177/1362361314539858

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*, 427–431. doi:10.1038/nature12715

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kang, S., Choi, J., Hong, S., & Lee, H. (2013). Up-Regulation of microRNA* Strands by Their Target Transcripts. *International Journal of Molecular Sciences*, *14*, 13231–13240. doi:10.3390/ijms140713231

Kenny, E. M., Cormican, P., Furlong, S., Heron, E., Kenny, G., Fahey, C., … Morris, D. W. (2014). Excess of rare novel loss-of-function variants in synaptic genes in schizophrenia and autism spectrum disorders. *Molecular Psychiatry*, *19*(8), 872–9. doi:10.1038/mp.2013.127

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, *39*(10), 1278–84. doi:10.1038/ng2135

Kim, J., Inoue, K., Ishii, J., Vanti, W. B., Voronov, S. V, Murchison, E., … Abeliovich, A. (2007). A MicroRNA feedback circuit in midbrain dopamine neurons. *Science*, *317*(5842), 1220–4. doi:10.1126/science.1140481

Kim, Y. S., Leventhal, B. L., Koh, Y.-J., Fombonne, E., Laska, E., Lim, E.-C., … Grinker, R. R. (2011). Prevalence of autism spectrum disorders in a total population sample. *The American Journal of Psychiatry*, *168*(9), 904–12. doi:10.1176/appi.ajp.2011.10101532

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., … Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, *2011*, bar030. doi:10.1093/database/bar030

Klei, L., Sanders, S. J., Murtha, M. T., Hus, V., Lowe, J. K., Willsey, a J., … Devlin, B. (2012). Common genetic variants, acting additively, are a major source of risk for autism. *Molecular Autism*, *3*(1), 9. doi:10.1186/2040-2392-3-9

Klintwall, L., Eldevik, S., & Eikeseth, S. (2013). Narrowing the gap: Effects of intervention on developmental trajectories in autism. *Autism*, (November). doi:10.1177/1362361313510067

Kong, A., Frigge, M., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., … Stefansson, K. (2012). Rate of de novo mutations, father's age, and disease risk. *Nature*, *488*(7412), 471–475. doi:10.1038/nature11396.Rate

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(Database issue), D152–7. doi:10.1093/nar/gkq1027

Kraemer, A., Chen, I., Henning, S., & Faust, A. (2013). UVA and UVB Irradiation Differentially Regulate microRNA Expression in Human Primary Keratinocytes. *PloS One*, *8*(12), 1–11. doi:10.1371/journal.pone.0083392

Krol, J., Loedige, I., & Filipowicz, W. (2010). The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics*, *11*(9), 597–610. doi:10.1038/nrg2843

Krüger, J., & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research*, *34*(Web Server issue), W451–4. doi:10.1093/nar/gkl243

Kye, M.-J., Liu, T., Levy, S. F., Xu, N. L., Groves, B. B., Bonneau, R., … Kosik, K. S. (2007). Somatodendritic microRNAs identified by laser capture and multiplex RT-PCR. *RNA*, *13*(8), 1224–34. doi:10.1261/rna.480407

Ladd-Acosta, C., Hansen, K. D., Briem, E., Kaufmann, W. E., & Feinberg, A. P. (2013). Common DNA methylation alterations in multiple brain regions in autism. *Molecular Psychiatry*, *19*(8), 862–871. doi:10.1038/mp.2013.114

Lai, M.-C., Lombardo, M. V, & Baron-Cohen, S. (2013). Autism. *Lancet*, *6736*(13). doi:10.1016/S0140-6736(13)61539-1

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., … Tuschl, T. (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, *129*, 1401–1414.

doi:10.1016/j.cell.2007.04.040

Landrigan, P. J. (2010). What causes autism? Exploring the environmental contribution. *Current Opinion in Pediatrics*, *22*(2), 219–25. doi:10.1097/MOP.0b013e328336eb9a

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Lappalainen, T., Sammeth, M., Friedländer, M. R., ’t Hoen, P. a C., Monlong, J., Rivas, M. a, … Ferreira, P. G. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–11. doi:10.1038/nature12531

Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V, Purcell, S. M., Perlis, R. H., … Witte, J. S. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, *45*(9), 984–94. doi:10.1038/ng.2711

Lee, S. Y., Sohn, K.-A., & Kim, J. H. (2012). MicroRNA-centric measurement improves functional enrichment analysis of co-expressed and differentially expressed microRNA clusters. *BMC Genomics*, *13*, S17. doi:10.1186/1471-2164-13-S7-S17

Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S. C., Frese, K., … Keller, A. (2013). A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biology*, *14*(7), R78. doi:10.1186/gb-2013-14-7-r78

Levy, F. (2007). Theories of autism. *The Australian and New Zealand Journal of Psychiatry*, *41*(11), 859–68. doi:10.1080/00048670701634937

Lewis, B. P., Shih, I., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, *115*(7), 787–98.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–9. doi:10.1093/bioinformatics/btp352

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., & Yang, J.-H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, *42*(Database issue), D92–7. doi:10.1093/nar/gkt1248

Li, S., Tsai, K., Pan, H., Jeng, Y., Ho, M., & Li, W. (2012). MicroRNA 3’ end nucleotide modification patterns and arm selection preference in liver tissues. *BMC Systems Biology*, *6*.

Liang, T., Liu, C., & Ye, Z. (2013). Deep Sequencing of Small RNA Repertoires in Mice Reveals Metabolic Disorders-Associated Hepatic miRNAs. *PloS One*, *8*(11), 1–9. doi:10.1371/journal.pone.0080774

Lim, E. T., Raychaudhuri, S., Sanders, S. J., Stevens, C., Sabo, A., MacArthur, D. G., … Daly, M. J. (2013). Rare complete knockouts in

humans: population distribution and significant role in autism spectrum disorders. *Neuron*, *77*(2), 235–42. doi:10.1016/j.neuron.2012.12.029

Liu, G., Ding, M., Chen, J., Huang, J., Wang, H., Jing, Q., & Shen, B. (2010). Computational analysis of microRNA function in heart development. *Acta Biochim Biophys Sin*, (August), 662–670. doi:10.1093/abbs/gmq072.Advance

Liu, L., Sabo, A., Neale, B. M., Nagaswamy, U., Stevens, C., Lim, E., … Roeder, K. (2013). Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genetics*, *9*(4), e1003443. doi:10.1371/journal.pgen.1003443

Lopez, J. P., Diallo, A., Cruceanu, C., Fiori, L. M., Laboissiere, S., Guillet, I., … Ernst, C. (2015). Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing. *BMC Medical Genomics*, *8*(35), 1–18. doi:10.1186/s12920-015-0109-x

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., Dilavore, P. C., … Rutter, M. (2000). 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*, *30*(3).

Lu, C., Meyers, B. C., & Green, P. J. (2007). Construction of small RNA cDNA libraries for deep sequencing. *Methods*, *43*, 110–117. doi:10.1016/j.ymeth.2007.05.002

Lu, T.-P., Lee, C.-Y., Tsai, M.-H., Chiu, Y.-C., Hsiao, C. K., Lai, L.-C., & Chuang, E. Y. (2012). miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets. *PloS One*, *7*(8), e42390. doi:10.1371/journal.pone.0042390

Luo, R., Sanders, S. J., Tian, Y., Voineagu, I., Huang, N., Chu, S. H., … Geschwind, D. H. (2012). Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *American Journal of Human Genetics*, *91*(1), 38–55. doi:10.1016/j.ajhg.2012.05.011

Ma, J., Jiang, Z., He, S., Liu, Y., Chen, L., Long, K., & Jin, L. (2013). Intrinsic Features in MicroRNA Transcriptomes Link Porcine Visceral Rather than Subcutaneous Adipose Tissues to Metabolic Risk. *PloS One*, *8*(11). doi:10.1371/journal.pone.0080041

Marco, A., & Griffiths-Jones, S. (2012). Detection of microRNAs in color space. *Bioinformatics*, *28*(3), 318–23. doi:10.1093/bioinformatics/btr686

Marco, A., Hui, J. H. L., Ronshaugen, M., & Griffiths-Jones, S. (2010). Functional shifts in insect microRNA evolution. *Genome Biology and Evolution*, *2*, 686–96. doi:10.1093/gbe/evq053

Marco, A., Macpherson, J. I., Ronshaugen, M., & Griffiths-Jones, S. (2012). MicroRNAs from the same precursor have different targeting properties. *Silence*, *3*(1), 8. doi:10.1186/1758-907X-3-8

Marco, A., Ninova, M., Ronshaugen, M., & Griffiths-Jones, S. (2013). Clusters of microRNAs emerge by new hairpins in existing transcripts. *Nucleic Acids Research*, *41*(16), 7745–52. doi:10.1093/nar/gkt534

Masliah-Planchon, J., Pasmant, E., Luscan, A., Laurendeau, I., Ortonne, N., Hivelin, M., … Bièche, I. (2013). MicroRNAome profiling in benign and malignant neurofibromatosis type 1-associated nerve sheath tumors: evidences of PTEN pathway alterations in early NF1 tumorigenesis. *BMC Genomics*, *14*, 473. doi:10.1186/1471-2164-14-473

McClimens, A. (2008). This is my truth, tell me yours: exploring the internal tensions within collaborative learning disability research. *British Journal of Learning Disabilities*, *36*, 271–276. doi:10.1111/j.1468-3156.2007.00485.x

McCreight, J. C., Schneider, S. E., Wilburn, D. B., & Swanson, W. J. (2017). Evolution of microRNA in primates. *PLoS One*, *12*(6), 1–22.

Meijer, H. A., Smith, E. M., & Bushell, M. (2014). Regulation of miRNA strand selection: follow the leader? *Biochemical Society Transactions*, *42*(4), 1135–1140. doi:10.1042/BST20140142

Merikangas, a K., Segurado, R., Heron, E. a, Anney, R. J. L., Paterson, a D., Cook, E. H., … Gallagher, L. (2014). The phenotypic manifestations of rare genic CNVs in autism spectrum disorder. *Molecular Psychiatry*, (April), 1–7. doi:10.1038/mp.2014.150

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, *11*(1), 31–46.

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., … Corominas, R. (2012). Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell*, *151*, 1431–1442. doi:10.1016/j.cell.2012.11.019

Milton, D. E. M. (2012). On the ontological status of autism: the “double empathy problem.” *Disability & Society*, *27*(6), 883–887. doi:10.1080/09687599.2012.710008

Mizuguchi, Y., Mishima, T., Yokomuro, S., Arima, Y., Kawahigashi, Y., Shigehara, K., … Takizawa, T. (2011). Sequencing and bioinformatics-based analyses of the microRNA transcriptome in hepatitis B-related hepatocellular carcinoma. *PloS One*, *6*(1), e15304. doi:10.1371/journal.pone.0015304

Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M., & Dermitzakis, E. T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genetics*, *7*(7), e1002144. doi:10.1371/journal.pgen.1002144

Mor, M., Nardone, S., Sams, D. S., & Elliott, E. (2015). Hypomethylation of miR-142 promoter and upregulation of microRNAs that target the oxytocin receptor gene in the autism prefrontal cortex. *Molecular Autism*, *6*(46), 1–11. doi:10.1186/s13229-015-0040-1

Munch, E., Harris, R., & Mohammad, M. (2013). Transcriptome profiling of

microRNA by Next-Gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk. *PloS One*, *8*(2), 1–13. doi:10.1371/journal.pone.0050564

Mundalil Vasu, M., Anitha, A., Thanseem, I., Suzuki, K., Yamada, K., Takahashi, T., … Mori, N. (2014). Serum microRNA profiles in children with autism. *Molecular autism*, *5*(1), 40. doi:10.1186/2040-2392-5-40

Nemlich, Y., Greenberg, E., Ortenberg, R., Besser, M. J., Barshack, I., Jacob-hirsch, J., … Markel, G. (2013). MicroRNA-mediated loss of ADAR1 in metastatic melanoma promotes tumor growth. *Journal of Clini*, *123*(6). doi:10.1172/JCI62980DS1

Nguyen, L. S., Lepleux, M., Makhlouf, M., Martin, C., Fregeac, J., Siquier-pernet, K., … Colleaux, L. (2016). Profiling olfactory stem cells from living patients identifies miRNAs relevant for autism pathophysiology. *Molecular Autism*, *7*(1), 1–13. doi:10.1186/s13229-015-0064-6

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, *12*(6), 443–51. doi:10.1038/nrg2986

Nind, M., & Vinha, H. (2012). Doing research inclusively: bridges to multiple possibilities in inclusive research. *British Journal of Learning Disabilities*, *42*, 102–109. doi:10.1111/bld.12013

Ninova, M., Ronshaugen, M., & Griffiths-jones, S. (2014). Conserved Temporal Patterns of MicroRNA Expression in Drosophila Support a Developmental Hourglass Model. *Genome Biology and Evolution*, *6*(9), 2459–2467. doi:10.1093/gbe/evu183

O'Roak, B., Vives, L., Fu, W., & Egertson, J. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, *338*(6114), 1619–1622. doi:10.1126/science.1227764.Multiplex

Okada, Y., Muramatsu, T., Suita, N., Kanai, M., Kawakami, E., Iotchkova, V., … Tanaka, T. (2016). Significant impact of miRNA-target gene networks on genetics of human complex traits. *Scientific Reports*, *6*, 22223. doi:10.1038/srep22223

Ortega, F. (2009). The Cerebral Subject and the Challenge of Neurodiversity. *BioSocieties*, *4*(4), 425–445.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. doi:10.1038/nrg2934

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., & Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Research*, *37*, 155–158. doi:10.1093/nar/gkn809

Parikshak, N. N., Luo, R., Zhang, A., Won, H., Lowe, J. K., Chandran, V., … Geschwind, D. H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell*,

155(5), 1008–21. doi:10.1016/j.cell.2013.10.031

Pasquinelli, A. E. (2012). MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, *13*(4), 271–282. doi:10.1038/nrg3162

Pekham, H., McLaughlin, S., Ni, J., Rhodes, M., Malek, J., McKernan, K., & Blanchard, A. (2007). SOLiD Sequencing and 2-Base Encoding. *American Society of Human Genetics Annual Meeting; Poster 2624*. Retrieved from http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_057810.pdf

Pellicano, E., & Burr, D. (2012). When the world becomes “too real”: a Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–10. doi:10.1016/j.tics.2012.08.009

Pellicano, E., Dinsmore, A., & Charman, T. (2014). What should autism research focus upon? Community views and priorities from the United Kingdom. *Autism*, *18*(7), 756–70. doi:10.1177/1362361314529627

Peñagarikano, O., Abrahams, B. S., Herman, E. I., Winden, K. D., Gdalyahu, A., Dong, H., … Geschwind, D. H. (2011). Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell*, *147*(1), 235–46. doi:10.1016/j.cell.2011.08.040

Peng, T., Lv, Q., Zhang, J., Li, J., Du, Y., & Zhao, Q. (2011). Differential expression of the microRNAs in superior and inferior spikelets in rice (Oryza sativa). *Journal of Experimental Botany*, *62*(14), 4943–54. doi:10.1093/jxb/err205

Peng, Z., Li, J., Li, Y., Yang, X., Feng, S., Han, S., & Li, J. (2013). Downregulation of miR-181b in mouse brain following ischemic stroke induces neuroprotection against ischemic injury through targeting heat shock protein A5 and ubiquitin carboxyl-terminal hydrolase isozyme L1. *Journal of Neuroscience Research*, *91*(10), 1349–62. doi:10.1002/jnr.23255

Pickles, A., Le Couteur, A., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., … Green, J. (2016). Parent-mediated social communication therapy for young children with autism (PACT): long-term follow-up of a randomised controlled trial. *The Lancet*, *6736*(16). doi:10.1016/S0140-6736(16)31229-6

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., … Wang, Z. (2014). Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *The American Journal of Human Genetics*, *94*, 677–94. doi:10.1016/j.ajhg.2014.03.018

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., … Abrahams, B. S. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, *466*(7304), 368–72. doi:10.1038/nature09146

Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable Identification of

Genomic Variants from RNA-Seq Data. *The American Journal of Human Genetics*, *93*(4), 641–651. doi:10.1016/j.ajhg.2013.08.008

Popov, N. T., Madjirova, N. P., & Minkov, I. N. (2012). Micro RNA HSA-486-3P Gene Expression Profiling in the Whole Blood of Patients with Autism. *Biotechnology & Biotechnological Equipment*, *26*(6). doi:10.5504/BBEQ.2012.0093

Pramparo, T., Pierce, K., Lombardo, M. V., Carter Barnes, C., Marinero, S., Ahrens-Barbeau, C., … Courchesne, E. (2015). Prediction of Autism by Translation and Immune/Inflammation Coexpressed Genes in Toddlers From Pediatric Community Practices. *JAMA Psychiatry*, *92093*, 1–9. doi:10.1001/jamapsychiatry.2014.3008

Rademakers, R., Eriksen, J. L., Baker, M., Robinson, T., Ahmed, Z., Lincoln, S. J., … Dickson, D. W. (2008). Common variation in the miR-659 binding-site of GRN is a major risk factor for TDP43-positive frontotemporal dementia. *Human Molecular Genetics*, *17*(23), 3631–3642. doi:10.1093/hmg/ddn257

Rajendran, G., & Mitchell, P. (2007). Cognitive theories of autism. *Developmental Review*, *27*(2), 224–260. doi:10.1016/j.dr.2007.02.001

Raponi, M., Dossey, L., Jatkoe, T., Wu, X., & Chen, G. (2009). MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. *Cancer Research*, *69*, 5776–5783. doi:10.1158/0008-5472.CAN-09-0587

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., & Hatzigeorgiou, A. G. (2012). Functional microRNA targets in protein coding sequences. *Bioinformatics*, *28*(6), 771–6. doi:10.1093/bioinformatics/bts043

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 2–3.

Ritchie, W., Flamant, S., & Rasko, J. E. J. (2009). Predicting microRNA targets and functions: traps for the unwary. *Nature Methods*, *6*(6), 397–8. doi:10.1038/nmeth0609-397

Robertson, S., & Ne'eman, A. (2008). Autistic Acceptance, the College Campus, and Technology: Growth of Neurodiversity in Society and Academia. *Disability Studies Quarterly*, *28*(4).

Robinson, E. B., Pourcain, B. S., Anttila, V., Kosmicki, J. A., Bulik-sullivan, B., Grove, J., … Daly, M. J. (2016). Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature Genetics*, *48*(5), 6–11. doi:10.1038/ng.3529

Romero-Cordoba, S., Rodriguez-Cuevas, S., Rebollar-Vega, R., Quintanar-Jurado, V., Maffuz-Aziz, A., Jimenez-Sanchez, G., … Hidalgo-Miranda, A. (2012). Identification and pathway analysis of microRNAs with no previous involvement in breast cancer. *PloS One*, *7*(3). doi:10.1371/journal.pone.0031904

Saba, R., & Schratt, G. M. (2010). MicroRNAs in neuronal development, function and dysfunction. *Brain Research*, *1338*, 3–13.

doi:10.1016/j.brainres.2010.03.107

Sanchez-Diaz, P., Hsiao, T., & Chang, J. (2013). De-regulated microRNAs in pediatric cancer stem cells target pathways involved in cell proliferation, cell cycle and development. *PloS One*, *8*(4). doi:10.1371/journal.pone.0061622

Sarachana, T., Zhou, R., Chen, G., Manji, H. K., & Hu, V. W. (2010). Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. *Genome Medicine*, *2*(4), 23.

Schmittgen, T. D., & Livak, K. J. (2008). Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, *3*(6), 1101–1108. doi:10.1038/nprot.2008.73

Schratt, G. M., Tuebing, F., Nigh, E. a, Kane, C. G., Sabatini, M. E., Kiebler, M., & Greenberg, M. E. (2006). A brain-specific microRNA regulates dendritic spine development. *Nature*, *439*(7074), 283–9. doi:10.1038/nature04367

Seo, J.-S., Ju, Y. S., Lee, W.-C., Shin, J.-Y., Lee, J. K., Bleazard, T., … Kim, Y. T. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Research*, *22*(11), 2109–19. doi:10.1101/gr.145144.112

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 668. doi:10.3389/fnhum.2013.00668

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., … Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, *498*(7453), 236–40. doi:10.1038/nature12172

Sherry, S. T., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., … Res, K. S. G. (2001). dbSNP : the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311.

Shi, B., Gao, W., & Wang, J. (2012). Sequence Fingerprints of MicroRNA Conservation. *PLoS One*, *7*(10), 1–6. doi:10.1371/journal.pone.0048256

Siegel, G., Obernosterer, G., Fiore, R., Oehmen, M., Bicker, S., Christensen, M., … Schratt, G. M. (2009). A functional screen implicates microRNA-138-dependent regulation of the depalmitoylation enzyme APT1 in dendritic spine morphogenesis. *Nature Cell Biology*, *11*(6), 705–16. doi:10.1038/ncb1876

Skuse, D., Warrington, R., Bishop, D., Chowdhury, U., Lau, J., Mandy, W., & Place, M. (2004). The developmental, dimensional and diagnostic interview (3di): a novel computerized assessment for autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*(5), 548–58. doi:10.1097/00004583-200405000-00008

Slotkin, W., & Nishikura, K. (2013). Adenosine-to-inosine RNA editing and human disease. *Genome Medicine*, *5*(105), 1–13.

Slowikowski, K., Hu, X., & Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues, and pathways affected by risk loci. *Bioinformatics*, 10–11.

Snow, A. V, Lecavalier, L., & Houts, C. (2009). The structure of the Autism Diagnostic Interview-Revised: diagnostic and phenotypic implications. *The Journal of Child Psychology and Psychiatry*, *6*(50), 734–742. doi:10.1111/j.1469-7610.2008.02018.x

Soh, J., Iqbal, J., & Queiroz, J. (2013). MicroRNA-30c reduces hyperlipidemia and atherosclerosis in mice by decreasing lipid synthesis and lipoprotein secretion. *Nature Medicine*, *19*(7). doi:10.1038/nm.3200

Sokolov, M., Panyutin, I., & Neumann, R. (2012). Unraveling the global microRNAome responses to ionizing radiation in human embryonic stem cells. *PloS One*, *7*(2). doi:10.1371/journal.pone.0031028

Somel, M., Liu, X., Tang, L., Yan, Z., Hu, H., Guo, S., … Khaitovich, P. (2011). MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. *PLoS Biology*, *9*(12), e1001214. doi:10.1371/journal.pbio.1001214

Spitzer, R. L., & Wakefield, J. C. (1999). DSM-IV diagnostic criterion for clinical significance: does it help solve the false positives problem? *The American Journal of Psychiatry*, *156*(12), 1856–64.

Stamova, B., Ander, B. P., Barger, N., Sharp, F. R., & Schumann, C. M. (2015). Specific Regional and Age-Related Small Noncoding RNA Expression Patterns Within Superior Temporal Gyrus of Typical Human Brains Are Less Distinct in Autism Brains. *Journal of Child Neurology*, *30*(14), 1930–1946. doi:10.1177/0883073815602067

Talebizadeh, Z., Butler, M. G., & Theodoro, M. F. (2008). Feasibility and relevance of examining lymphoblastoid cell lines to study role of microRNAs in autism. *Autism Research*, *1*(4), 240–50. doi:10.1002/aur.33

Tan, G. C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, M., … Dibb, N. J. (2014). 5’ isomiR variation is of functional and evolutionary importance. *Nucleic Acids Research*, *42*(14), 9424–9435. doi:10.1093/nar/gku656

Tanic, M., & Andrés, E. (2013). MicroRNA-based molecular classification of non-BRCA1/2 hereditary breast tumours. *British Journal of Cancer*, *109*, 2724–2734. doi:10.1038/bjc.2013.612

Tatro, E., Risbrough, V., Soontornniyomkij, B., Young, J., Shumaker-Armstrong, S., Jeste, D., & Achim, C. (2013). Short-Term Recognition Memory Correlates with Regional CNS Expression of microRNA-138 in Mice. *The American Journal of Geriatric Psychiatry*, *21*(5), 461–473.

Tong, Y., Peng, H., Zhan, C., Fan, L., Ai, T., & Wang, S. (2013). Genome-wide analysis reveals diversity of rice intronic miRNAs in sequence structure, biogenesis and function. *PloS One*, *8*(5), e63938.

doi:10.1371/journal.pone.0063938

Tsai, K., Leung, C., Lo, Y., Chen, T., & Chan, W. (2016). Arm Selection Preference of MicroRNA-193a Varies in Breast Cancer. *Scientific Reports*, *6*, 1–13. doi:10.1038/srep28176

Vaishnavi, V., Manikandan, M., Tiwary, B. K., & Munirajan, A. K. (2013). Insights on the Functional Impact of MicroRNAs Present in Autism-Associated Copy Number Variants. *PLoS One*, *8*(2), e56781. doi:10.1371/journal.pone.0056781

Vargas, D. L., Nascimbene, C., Krishnan, C., Zimmerman, A. W., & Pardo, C. a. (2005). Neuroglial activation and neuroinflammation in the brain of patients with autism. *Annals of Neurology*, *57*(1), 67–81. doi:10.1002/ana.20315

Verhoeff, B. (2013). The autism puzzle: challenging a mechanistic model on conceptual and historical grounds. *Philosophy, Ethics, and Humanities in Medicine*, *8*(1), 17. doi:10.1186/1747-5341-8-17

Visvanathan, J., Lee, S., Lee, B., Lee, J. W., & Lee, S.-K. (2007). The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. *Genes & Development*, *21*(7), 744–9. doi:10.1101/gad.1519107

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., … Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, *474*(7351), 380–4. doi:10.1038/nature10110

Walsh, P., Elsabbagh, M., Bolton, P., & Singh, I. (2011). In search of biomarkers for autism: scientific, social and ethical challenges. *Nature Reviews Neuroscience*, *12*, 603–612.

Wang, F., Sun, G., Zou, Y., Fan, L., & Song, B. (2012). Lack of association of miR-146a rs2910164 polymorphism with gastrointestinal cancers: evidence from 10206 subjects. *PloS One*, *7*(6), e39623. doi:10.1371/journal.pone.0039623

Wang, G., Walt, J. M. Van Der, Mayhew, G., Li, Y., Zu, S., Scott, W. K., … Vance, J. M. (2008). Variation in the miRNA-433 Binding Site of FGF20 Confers Risk for Parkinson Disease by Overexpression of a -Synuclein. *The American Journal of Human Genetics*, *82*, 283–289. doi:10.1016/j.ajhg.2007.09.021.

Wang, T., Guo, H., Xiong, B., Stessman, H. a F., Wu, H., Coe, B. P., … Eichler, E. E. (2016). De novo genic mutations among a Chinese autism spectrum disorder cohort. *Nature Communications*, *7*(13), 13316. doi:10.1038/ncomms13316

Wei, L. Q., Yan, L. F., & Wang, T. (2011). Deep sequencing on genome-wide scale reveals the unique composition and expression patterns of microRNAs in developing pollen of Oryza sativa. *Genome Biology*, *12*(6), R53. doi:10.1186/gb-2011-12-6-r53

Willsey, A. J., Sanders, S. J., Li, M., Dong, S., Tebbenkamp, A. T., Muhle, R. A., … State, M. W. (2013). Coexpression Networks Implicate

Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism. *Cell*, *155*(5), 997–1007. doi:10.1016/j.cell.2013.10.020

Wright, C., Turner, J. a, Calhoun, V. D., & Perrone-Bizzozero, N. (2013). Potential Impact of miR-137 and Its Targets in Schizophrenia. *Frontiers in Genetics*, *4*, 58. doi:10.3389/fgene.2013.00058

Wu, C., Lin, H., Wang, Q., Chen, W., Luo, H., Chen, W., & Zhang, H. (2012). Discrepant expression of microRNAs in transparent and cataractous human lenses. *Investigative Ophthalmology & Visual Science*, *53*(7), 3906–12. doi:10.1167/iovs.11-9178

Wu, X., Shao, X., Meng, X., Zhang, X., Zhu, L., Liu, S., … Xiao, H. (2011). Genome-wide analysis of microRNA and mRNA expression signatures in hydroxycamptothecin-resistant gastric cancer cells. *Acta Pharmacologica Sinica*, *32*(2), 259–69. doi:10.1038/aps.2010.204

Wu, Y. E., Parikshak, N. N., Belgard, T. G., & Geschwind, D. H. (2016). Genome-wide, integrative analysis implicates microRNA dysregulation in autism spectrum disorder. *Nature Neuroscience*, *19*(11). doi:10.1038/nn.4373

Wu, Y.-Q., Chen, D.-J., He, H.-B., Chen, D.-S., Chen, L.-L., Chen, H.-C., & Liu, Z.-F. (2012). Pseudorabies virus infected porcine epithelial cell line generates a diverse set of host microRNAs and a special cluster of viral microRNAs. *PloS One*, *7*(1), e30988. doi:10.1371/journal.pone.0030988

Xie, F., Stewart, C. N., Taki, F. a, He, Q., Liu, H., & Zhang, B. (2013). High-throughput deep sequencing shows that microRNAs play important roles in switchgrass responses to drought and salinity stress. *Plant Biotechnology Journal*, *159*, 1–13. doi:10.1111/pbi.12142

Xu, B., Karayiorgou, M., & Gogos, J. (2010). MicroRNAs in psychiatric and neurodevelopmental disorders. *Brain Research*, *1338C*, 78–88. doi:10.1016/j.brainres.2010.03.109.microRNAs

Xu, W., Xu, J., Liu, S., Chen, B., Wang, X., Li, Y., … Wu, J. (2011). Effects of common polymorphisms rs11614913 in miR-196a2 and rs2910164 in miR-146a on cancer susceptibility: a meta-analysis. *PloS One*, *6*(5), e20471. doi:10.1371/journal.pone.0020471

Yan, Q., Li, W., Tang, Q., Yao, S., Lv, Z., Feng, N., … Lu, C. (2013). Cellular microRNAs 498 and 320d regulate herpes simplex virus 1 induction of Kaposi's sarcoma-associated herpesvirus lytic replication by targeting RTA. *PloS One*, *8*(2), e55832. doi:10.1371/journal.pone.0055832

Yan, Z., Xiong, Y., Xu, W., Gao, J., Cheng, Y., Wang, Z., … Zheng, G. (2012). Identification of hsa-miR-335 as a prognostic signature in gastric cancer. *PloS One*, *7*(7), e40037. doi:10.1371/journal.pone.0040037

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7),

565–9. doi:10.1038/ng.608

Yi, R., Zhu, Z., Hu, J., Qian, Q., Dai, J., & Ding, Y. (2013). Identification and expression analysis of microRNAs at the grain filling stage in rice(Oryza sativa L.)via deep sequencing. *PloS One*, *8*(3), e57863. doi:10.1371/journal.pone.0057863

Yuen, R. K. C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., … Scherer, S. W. (2016). Genome-wide characteristics of de novo mutations in autism. *NPJ Genomic Medicine*, *1*. doi:10.1038/npjgenmed.2016.27

Yuen, R. K. C., Thiruvahindrapuram, B., Merico, D., Walker, S., Tammimies, K., Hoang, N., … Scherer, S. W. (2015). Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature Medicine*, (October 2014). doi:10.1038/nm.3792

Yunta, M., Nieto-Díaz, M., Esteban, F. J., Caballero-López, M., Navarro-Ruíz, R., Reigada, D., … Maza, R. M. (2012). MicroRNA dysregulation in the spinal cord following traumatic injury. *PloS One*, *7*(4), e34534. doi:10.1371/journal.pone.0034534

Zhao, X., Leotta, A., Kustanovich, V., Lajonchere, C., Geschwind, D. H., Law, K., … Wigler, M. (2007). A unified genetic theory for sporadic and inherited autism. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(31), 12831–6. doi:10.1073/pnas.0705803104

Zhu, S., Pan, W., Song, X., Liu, Y., Shao, X., Tang, Y., … Qian, Y. (2012). The microRNA miR-23b suppresses IL-17-associated autoimmune inflammation by targeting TAB2, TAB3 and IKK-α. *Nature Medicine*, *18*(7), 1077–86. doi:10.1038/nm.2815

Ziats, M. N., & Rennert, O. M. (2013). Identification of differentially expressed microRNAs across the developing human brain. *Molecular Psychiatry*, (May), 1–5. doi:10.1038/mp.2013.93

Zou, M., Li, D., Lv, R., Zhou, Y., Wang, T., Liu, J., … Wang, L. (2012). Association between two single nucleotide polymorphisms at corresponding microRNA and schizophrenia in a Chinese population. *Molecular Biology Reports*, *39*(4), 3385–91. doi:10.1007/s11033-011-1109-3