Title
Investigation of sex-differential genetic risk factors for autism spectrum disorders

Author
Werling, Donna Marie

Publication Date
2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Investigation of sex-differential genetic risk factors for autism spectrum disorders

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Neuroscience

by

Donna Marie Werling

2014

ABSTRACT OF THE DISSERTATION

Investigation of sex-differential genetic risk factors for autism spectrum disorders

by

Donna Marie Werling

Doctor of Philosophy in Neuroscience

University of California, Los Angeles, 2014

Professor Daniel H. Geschwind, Chair

Autism spectrum disorders (ASDs) are pervasive neurodevelopmental disorders that affect more males than females, and the mechanisms responsible for increasing males' risk or protecting females are not understood. This sex biased prevalence is consistent across time and populations, suggesting that an understanding of the processes driving sex-differential risk would likely be informative of fundamental pathophysiology in ASD. One known component of ASD risk is genetic variation. Thus, here I apply several approaches that leverage current knowledge of ASD genetics to investigate the role and mechanisms of sex-differential biology in ASD risk. First, I evaluate a cohort of families with more than one autistic child for evidence of sex-differential, familial risk variation. Second, I use genetic linkage analysis to identify sex-differential risk loci in families from the same multiplex cohort. Third, I characterize gene expression patterns in typical human neocortex to identify points of interaction between typical sexual dimorphism and genes known to carry risk variants for ASD.

I find that recurrence rates for ASD diagnoses in multiplex families are consistent with a female protective model, in which females require more deleterious genetic variation to be affected with ASD and this greater genetic load is shared with females' siblings. I also identify several chromosomal loci with evidence of genetic linkage in families either with (chromosome 8p21.2 and 8p12), or without (chromosome 1p31.3), an autistic female. No significant common variants are found in either region that can account for this linkage; these loci will be further investigated by targeted sequencing to identify rare risk variants. Gene expression analyses show that known ASD risk genes are not differentially expressed in males or females in the prenatal or adult human neocortex. However, astrocyte markers and gene sets implicated in immune function and inflammatory processes are expressed at higher levels in males. This suggests that sex-differential factors may operate downstream from, or interact with, ASD risk genes, as opposed to directly regulating the expression of these genes. Overall, findings from these multiple approaches provide valuable context for the function of sex-differential biology in ASD etiology, and suggest promising directions for future research.

The dissertation of Donna Marie Werling is approved.

Arthur P. Arnold

Rita M. Cantor

Nelson B. Freimer

Daniel H. Geschwind, Committee Chair

University of California, Los Angeles

2014

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGEMENTS

Chapter 1 is a version of, “Sex differences in autism spectrum disorders” published April 2013 in *Current Opinion in Neurology*, Volume 26, Issue 2, p 146-53. DOI: 10.1097/WCO.0b013e32835ee548. Wolters Kluwer Health Lippincott Williams & Wilkins©. This work was co-authored by Daniel H. Geschwind, principal investigator and committee chair, who has agreed to the publication of this review in this dissertation.

Chapter 3 is a version of “Replication of linkage at chromosome 20p13 and identification of suggestive sex-differential risk loci for autism spectrum disorder” published February 17, 2014 in *Molecular Autism*, Volume 5, Issue 1, p 13. DOI: 10.1186/2040-2392-5-13. This work was co-authored by Jennifer K. Lowe, Rui Luo, Rita M. Cantor, and Daniel H. Geschwind. I, Donna M. Werling, carried out all subject selection, linkage and association analyses, and prepared and drafted the manuscript. Jennifer K. Lowe performed preprocessing and quality control on genotype data and provided assistance with genetic analyses. Rui Luo carried out the genotype imputation for all samples. Rita M. Cantor advised on the design and interpretation of the study and manuscript text. Daniel H. Geschwind conceived of the study, participated in its design and interpretation, and assisted with manuscript preparation. All co-authors have agreed to the re-publication of this work in this dissertation.

Chapter 4 describes work co-authored by Neelroop N. Parikshak and Daniel H. Geschwind. I, Donna M. Werling, conceived of the study design, processed input data sets, and performed all analyses. Neelroop Parikshak directed the collection and pre-processing of RNA-sequencing data from the adult cortex replication set, advised on analytic methods, and wrote the base code for several functions applied during pre-processing and analysis. Daniel H. Geschwind

advised on study design and interpretation and assisted with manuscript preparation. All co-authors have agreed to the publication of this study in this dissertation.

I also gratefully acknowledge sources of financial support for this work. Studies presented in chapters 2-4 of this dissertation were supported by an Autism Center for Excellence (ACE) Network grant 5R01 R01MH081754 and ACE Center grant 3P50 HD055784 awarded to Daniel H Geschwind. The Autism Genetics Resource Exchange, also utilized in chapters 2-4, is a program of Autism Speaks and is supported by grant NIMH 1U24MH081810 awarded to Clara M. Lajonchere. I, Donna M. Werling, was supported by the Chancellor's Prize from the UCLA Graduate Division, a fellowship from the Achievement Rewards for College Scientists (ARCS) Foundation, and a Ruth L. Kirschstein National Research Service Award from the National Institute of Mental Health, F31-MH093086.

VITA

Donna Marie Werling

Education:

Bachelor of Science in Psychology, Duke University, Durham, NC, 2008.

Publications:

2014 Lowe, J.K., Werling, D.M., Constantino, J.N., Cantor, R.M., & Geschwind, D.H. (2014). Quantitative linkage analysis to the autism endophenotype social responsiveness identifies genome-wide significant linkage to two regions on chromosome 8. *American Journal of Psychiatry*, in press.

2014 Werling, D.M., Lowe, J.K., Luo, R., Cantor, R.M., & Geschwind, D.H. (2014). Replication of linkage for autism spectrum disorder at chromosome 20p13 and identification of suggestive sex-differential risk loci. *Molecular Autism, 5*(1), 13.

2013 Werling, D.M., & Geschwind, D. (2013). Understanding sex bias in autism spectrum disorder. *Proceedings of the National Academy of Sciences, 110*(13), 4868-4869.

2013 Werling, D.M., & Geschwind, D. (2013). Sex differences in autism spectrum disorders. *Current Opinion in Neurology, 26*(2), 146-153.

2010 Watson, K., Werling, D.M., Zucker, N., & Platt, M. (2010). Altered social reward and attention in anorexia nervosa. *Frontiers in Psychopathology, 1*, 1:7.

Presentations:

2014 Werling, D.M., Parikshak, N.N., & Geschwind, D.H. Sex-differential gene expression in human brain: Implications for autism spectrum disorders. *International Meeting for Autism Research.* Atlanta, GA.

2013 Werling, D.M. Intersection of sex-differential and ASD-specific gene expression patterns in human brain. *20th Annual Molecular Psychiatry Meeting*. Park City, UT.

2010 Berg*, J.M., & Werling*, D.M. Introduction to Autism and Fragile X Syndrome: Uncovering the biological underpinnings using genetically-engineered mouse models. *Achievement Rewards for College Scientists private donor event.* Los Angeles, CA.

*Joint oral presentation.

2009 Werling, D.M. & Oguro-Ando, A. Behavioral and morphological analysis of cytoplasmic FMRP-interacting protein 1 (Cyfip1), an autism candidate gene. *UCLA Intellectual and Developmental Disabilities Research Center Annual Retreat.* Los Angeles, CA.

Grants awarded:

2011-2014 Ruth L. Kirschstein NRSA from NIMH, F31-MH093086

2008 Trinity College Research Forum in Neuroscience (Duke University)

2007 Undergraduate Research Support Grant, (Trinity College Undergraduate Research Support Office, Duke University)

Honors:

2008-2012 Achievement Rewards for College Scientists (ARCS) Foundation Scholar

2008-2009 Chancellor's Prize, UCLA

2008 Summa cum laude, Duke University

2008 Graduation with Distinction (Honors Thesis), Duke University

2007 Inducted into Phi Beta Kappa, academic honor society

2004-2008 Dean's List with Distinction, Duke University

CHAPTER 1:

Sex differences in autism spectrum disorders

1.1: Introduction

Sexually dimorphic disease prevalence is well recognized, but poorly understood. For example, many disorders with autoimmune etiologies, such as multiple sclerosis and systemic lupus erythematosis, are female predominant (Whitacre, 2001), whereas some neurodevelopmental disorders, such as attention deficit hyperactivity disorder and language impairment show a male bias (Barbaresi et al., 2002; Szatmari, Offord, & Boyle, 1989; Viding et al., 2004). Autism spectrum disorders (ASDs) are prototypical in this regard, as they show a striking male bias in prevalence, with approximately 4 affected males for every 1 affected female. The consistency of this observation across time and populations strongly implicates the involvement of sex-specific biological factors in ASD etiology. However, we have yet to definitively identify the underlying mechanism through which these pathways interact to give rise to the male preponderance among individuals with ASDs. In recent years, increased priority has been placed on the inclusion and study of autistic females, while geneticists have made considerable headway in identifying novel genetic risk variants for ASD, putting us now in a position to assess relationships between genetic risk factors, hormones, and observed patterns of sex-differential vulnerability to ASDs. Toward this goal, here we review patterns of sex bias in ASD prevalence and phenotypic presentation, and evaluate the evidence for several hypotheses that could explain the biological basis of the male bias in ASD. We also identify areas of research where additional work is needed to advance our understanding of the interactions between sex-differential biology and risk factors for ASD.

1.2: ASD prevalence in males and females

From the first published descriptions of autism, it has been a male-typical disorder: 8 of the 11 cases described by Kanner (Kanner, 1943), and all 4 cases described by Asperger (Asperger, 1944), were male. Prevalence surveys conducted since have reported a range of male biases from 1.33:1 male:female (M:F) to 15.7:1 (Fombonne, 2009), and a commonly referenced consensus ratio of ~4:1. Intelligence level affects this sex ratio: males are substantially over-represented among high-functioning cases, and males and females are more equally represented among cases with severe intellectual disability (ID) (Banach et al., 2009; Fombonne, 1999; Volkmar, Szatmari, & Sparrow, 1993; Yeargin-Allsopp et al., 2003); a 1999 review reported median sex ratios of 6:1 among normal-functioning subjects and 1.7:1 among cases with moderate to severe ID (Fombonne, 1999).

Several biological factors could explain this relationship between IQ and the sex ratio, but it should also be recognized that this could also, at least partially, reflect ascertainment bias. Co-morbid ID increases females' likelihood of acquiring an ASD diagnosis, and conversely high-functioning females may go undiagnosed. The wide variation in the sex ratio reported by different sites in the Autism and Developmental Disabilities Monitoring (ADDM) Network and findings from large-scale population screening for ASD in a South Korean community where clinically ascertained samples show higher M:F ratios than less biased population screening are consistent with this hypothesis (Developmental Disabilities Monitoring Network, 2012; Y. S. Kim et al., 2011). Generally, high sex ratios have been found by studies that predominantly identified subjects via treatment facilities or disability registries (Lord & Schopler, 1985; Volkmar et al., 1993; Yeargin-Allsopp et al., 2003), including more recent studies of records from Boston area hospital records (Kohane et al., 2012) and Taiwanese disability registries (Lai,

Tseng, Hou, & Guo, 2012). In contrast, sex ratios between 1.7:1 in high-functioning ASD cases, and 2.3:1 in cases with ASD and co-morbid ID were found in an epidemiological, population-screening study for ASDs in Finland (Mattila et al., 2011), although the same trend was not found in England, where a 9 to 1 M:F ratio was observed in high functioning individuals with an ASD (Brugha et al., 2011).

Additionally, the degree of male bias in ASD prevalence appears to be dependent on family structure. Analysis of families from the Interactive Autism Network (IAN) found a sex ratio of 5.8:1 for probands from 968 simplex families, and a sex ratio of 3.27:1 for 134 multiplex families (Constantino, Zhang, Frazier, Abbacchi, & Law, 2010). Two surveillance studies of high-risk infant siblings of autistic probands also report attenuated male biases (2.8:1 (Ozonoff et al., 2011) and 1.65:1 (Zwaigenbaum et al., 2012)) among ASD-diagnosed siblings in these inherently multiplex families. While the risk architectures in simplex and multiplex families may interact differently with sex-specific biology, these observations may instead be indicative of another ascertainment issue wherein parents with an affected child are more likely to notice ASD symptoms and pursue diagnoses for their subsequent female children regardless of her intellectual functioning. Other parents without older affected children may not seek diagnoses for their high-functioning daughters' condition, but these children would likely be identified in community screens, reducing the observed ratio of males to females as suggested above.

Overall, prevalence studies demonstrate that ASD is consistently over-represented in males as compared to females. But, we currently do not understand the extent of this over-representation, or the degree to which this male biased prevalence is related to intellectual functioning or ascertainment methods in addition to the influence of sex-differential genetic or hormonal factors.

1.3: Presentation of ASD symptoms and related phenotypes in males and females

In contrast with the higher proportion of diagnosed autistic females than males with ID, many studies find no sex differences in overall composite ASD severity as measured on several standard assessment tools (Carter et al., 2007; Kopp & Gillberg, 2011; Lai et al., 2012; Lai et al., 2011; Mayes & Calhoun, 2011; Zwaigenbaum et al., 2012). This suggests that among those who meet diagnostic criteria, females are not more severely affected. However, differences emerge when each core symptom domain of ASD is considered separately, and sex differences observed in cases tend to reflect sex differences observed in the typical population. Males with ASD are found to show more externalizing behavior problems than females, such as aggressive behavior, hyperactivity, reduced prosocial behavior, and increased repetitive/restricted behaviors and interests (Bolte, Duketis, Poustka, & Holtmann, 2011; Giarelli et al., 2010; Hattier, Matson, Tureck, & Horovitz, 2011; Mandy et al., 2012; Szatmari et al., 2012). Females with ASD show greater internalizing symptoms than boys, including anxiety, depression, and other emotional symptoms as reported by parents (Mandy et al., 2012; Solomon, Miller, Taylor, Hinshaw, & Carter, 2012); parents also more frequently endorse the item "avoids demands" for female cases on the Autism Spectrum Screening Questionnaire (ASSQ) (Kopp & Gillberg, 2011), perhaps reflecting girls' tendency to misbehave passively, as opposed to acting out. The observed sex differences raise the possibility that male-typical externalizing behaviors are more disruptive in the home or school setting than female-typical internalizing behaviors, preferentially prompting evaluation and diagnosis for boys, especially as compared to high-functioning girls. For girls then, ID may be more likely the secondary issue prompting evaluation and diagnosis. This scenario further implies that some proportion of the sex difference in ASD prevalence is attributable to biases inherent in the diagnostic process.

A recent study from the UK addressed this potential diagnosis gap by characterizing children with high autistic traits who met or fell short of the threshold for ASD diagnosis (Dworzynski, Ronald, Bolton, & Happe, 2012). A significantly smaller proportion of high-scoring girls met full ASD diagnostic criteria than males (38% versus 56%) whereas ASD-diagnosed girls had a higher mean total problem score (hyperactivity, anxiety, and conduct, peer, and prosocial problems) and a higher frequency of low IQ than ASD-diagnosed boys. Girls without diagnoses showed increased communication difficulties, but reduced social impairments as compared to non-diagnosed boys. Thus, it may be that relatively higher levels of social ability in females preclude full diagnosis of ASD, particularly for those who are high-functioning.

Phenotypic sex differences also become apparent when ASD traits are examined at different stages of development. In a study of children and adults with high-functioning ASD, scores from the ADI-R, which assesses lifetime incidence of ASD traits, were no different between males and females, but scores from the ADOS, which assesses current behavior, in autistic adults showed significantly less socio-communication impairment in females than males (Lai et al., 2011). These nuances potentially suggest that both sexes may be comparably affected early in life, but by adulthood, females display fewer socio-communication impairments in interactions with a clinician (ADOS). Whether this is a true remittance of a primary, biological deficit in social ability is unknown. Instead, it has been proposed that high-functioning females may be more likely than males to learn and apply social skills that mask, or camouflage, their difficulties with social interactions (Attwood, 2006; Lai et al., 2011), a phenomenon which may also affect the observed prevalence of ASD in females.

Nevertheless, whether the male-skewed prevalence of ASD is due to biased diagnosis of sex-differential presentations of the disease or to true sex differences in prevalence (or both),

sex-specific biology is likely to play a role. For the remainder of this chapter, we discuss the relationships between ASD and the two major drivers of sex-specific biology: genetics and hormones.

1.4: Sex differences in genetic contributions to ASD risk

Biological theories for the sex difference in ASD prevalence most frequently take the form of a multiple-threshold multifactorial liability model (Reich, Cloninger, & Guze, 1975), in which females have a higher threshold for reaching affection status than males (Figure 1.1A). Thus, genetic studies operating under this model hypothesize that females with ASD are likely to be carrying a higher heritable mutational "load" than affected males. This model predicts that relatives of female probands should be at increased risk for ASD as compared with relatives of male probands. This hypothesis is supported by a twin study finding higher risk to male co-twins of female probands than to female co-twins of male probands (Hallmayer et al., 2011). In contrast, other studies have failed to support the genetic loading hypothesis, including a study of two samples totaling 882 families (Goin-Kochel, Abbacchi, & Constantino, 2007) and a study of high risk siblings of autistic probands that found that only the sex of the sibling was a significant predictor of their future ASD status (Ozonoff et al., 2011). However, we note that these study designs inherently include children from both simplex and multiplex families, therefore calculating recurrence rates from a combination of children who carry *de novo* genetic risk variants (not shared with siblings) and children who carry heritable risk variants (likely shared with siblings).

More recently, a study of quantitative measures of autistic traits in more than 9000 dizygotic twin pairs from population-based cohorts provides the most conclusive demonstration

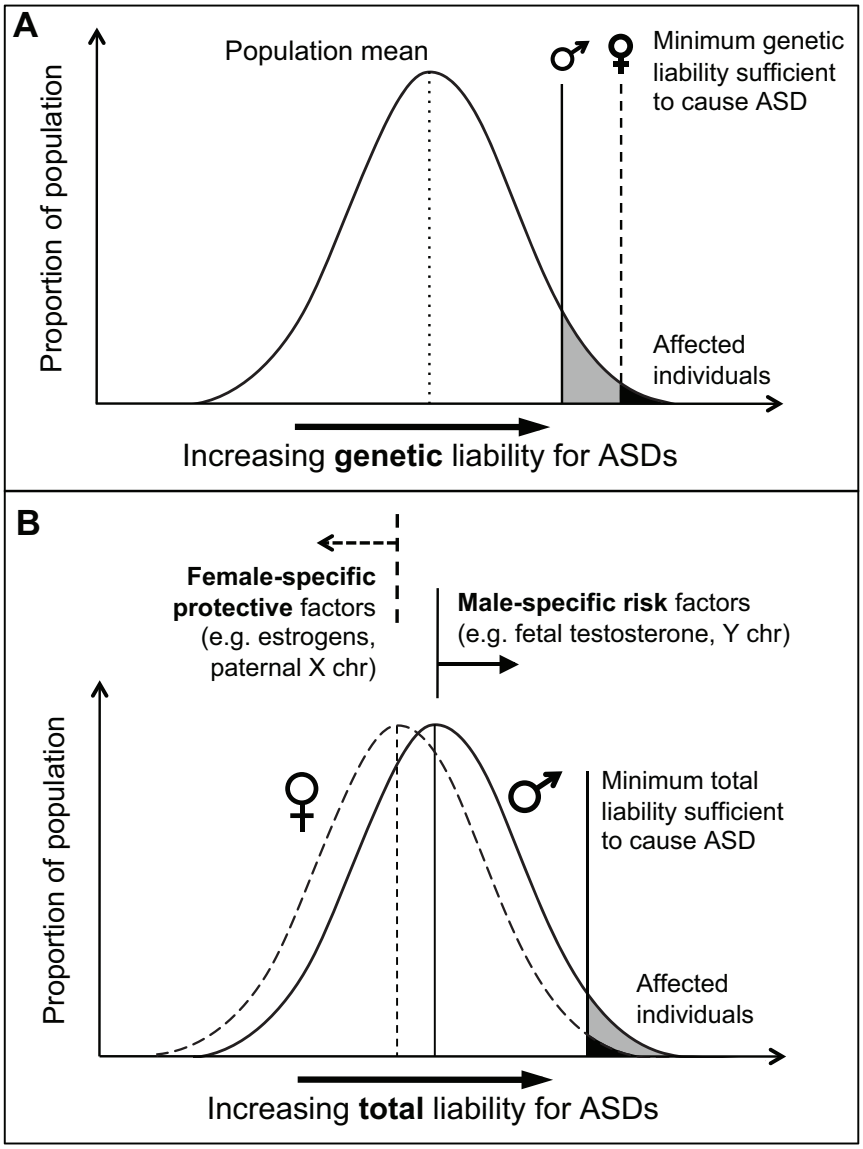


Figure 1.1: Multifactorial liability models for ASDs.
A) Multiple-threshold model in which genetic liability for ASD is normally distributed in the population and the minimum genetic liability sufficient to cause ASD (liability threshold) in females is greater than in males. B) Multifactorial liability model in which total liability for ASD, including contributions from genetic variation, environment, and other biological factors, is distributed in the population; female-specific factors shift females' total liability distribution away from, and male-specific factors shift males' distribution toward, a single threshold. Figure adapted from Reich et al. (1975).

of the female-protective model to date, showing that siblings of autistic females exhibit significantly greater autistic impairments than siblings of autistic males in both cohorts

(Robinson, Lichtenstein, Anckarsater, Happe, & Ronald, 2013). In contrast with recurrence rate studies for diagnosis, a binary variable, the use of a quantitative measure here allowed Robinson and colleagues to select the most severely impacted children as probands and it allowed for children with elevated but sub-threshold autistic behaviors to be included in the analysis. Broadly, this finding also supports a role for heritable variation in ASD liability under the threshold model.

There is additional experimental evidence for heritable loci with sex-differential penetrance. In one approach, multiplex family samples are divided into two groups for analysis: those with only affected male children ("male-only"), and those with at least one affected female child ("female-containing") to identify sex-differential genetic variation at several loci in male-only and female-containing families (Cantor et al., 2005; Lamb et al., 2005; Schellenberg et al., 2006; Stone et al., 2004; Szatmari et al., 2007; Weiss, Arking, Daly, & Chakravarti, 2009). However, only the male-only linkage signal at 17q21 has been successfully replicated, and the exact risk genes or variants responsible for these linkage peaks remain unknown. Other approaches have identified more defined sex related risk loci. For example, Lu and Cantor used a case-pseudocontrol genome-wide association test with sex as a factor to find two genome-wide significant single nucleotide polymorphisms (SNPs) within genes RYR2 (Ryanodine receptor 2) and UPP2 (uridine phosphorylase 2) (Lu & Cantor, 2012). Also, a study of rare copy number variants (CNVs) in ASD identified the first inherited autosomal variant with clear male-biased penetrance: males carrying a microdeletion in SHANK1 had high-functioning autism, while female relatives carrying the same microdeletion showed anxiety but did not meet diagnostic criteria for ASD (Sato et al., 2012). Subsequently, it was also reported that CNVs at chromosome 16p13.11 are enriched among male, but not female, cases with neurodevelopmental conditions,

including ASDs and speech and language delay (Tropeano et al., 2013). Although they cannot fully explain the male bias in ASD, these sex-differential linkage peaks, SNPs, the SHANK1 microdeletion, and 16p13.11 CNVs represent promising starting points for further work to elucidate the mechanism by which these inherited variants confer sex-differential ASD risk.

Aside from heritable variation, it is also plausible that some of females' hypothesized higher genetic load is caused by de novo variation (Berg & Geschwind, 2012) and is therefore not shared with relatives. In fact, close to 90% of ASD families have only one affected member (Constantino et al., 2010), suggesting that de novo variants of large effect may contribute significantly to ASD liability. This is supported by studies of chromosomal structural variation that show trends toward a higher proportion of female cases carrying a de novo CNV than male cases, as well as toward higher numbers of disrupted genes in females' CNVs than males' (Levy et al., 2011; Sanders et al., 2011; Sebat et al., 2007). One of these studies also observed a trend towards increased de novo CNV rate in unaffected female siblings (Sanders et al., 2011), consistent with the hypothesis that females can withstand more significant mutations than males before being affected with ASD. An elevated rate of de novo single nucleotide variants (SNVs) is also observed in exome sequences from autistic females (Neale et al., 2012; O'Roak et al., 2012), especially the most deleterious SNVs (nonsense, splice site, some missense) (Iossifov et al., 2012; Sanders et al., 2012), though again, these differences in SNV rate did not reach statistical significance. However, a newer study analyzing reported CNVs and SNVs in the Simons Simplex Collection and a large sample of children with neurodevelopmental disorders referred for clinical genetic evaluation did observe significant increases in the proportion of affected females carrying deleterious *de novo* and inherited variants, as well as a trend toward maternal transmission of inherited deleterious variants (Jacquemont et al., 2014). Taken together

with studies of familial recurrence rate, it appears that regardless of whether risk variants are inherited or de novo, in the face of a comparable degree of genetic liability, males are at increased risk for and females are protected from manifesting ASD symptoms that meet diagnostic criteria.

Sex chromosomal genes have been proposed to be key players in molecular mechanisms driving females' protection from ASD liability conferred by specific risk loci and/or by genome-wide mutational load (Figure 1.1B). An early theory proposed that ASD might be an X-linked disorder, in which females are protected from deleterious effects of X chromosomal mutations by compensatory transcription from their intact, second X chromosome. However, ASD transmission in most families does not follow an X-linked pattern, and while several ASD risk genes have been identified on the X chromosome (e.g. FMRP, MECP2, NLGN3, NLGN4X), all cause significant ID, indicating a more general role for X chromosome gene dosage in neural development.

Although ASD may not be X-linked in the Mendelian sense, sex chromosome complement may still modulate ASD risk. Sex chromosome aneuploidies provide test cases for this hypothesis, with an increased rate of ASD diagnosis in Turner syndrome (TS, XO, ~3% ASD) (Creswell & Skuse, 1999; Donnelly et al., 2000; Skuse et al., 1997), Klinefelter syndrome (KS, XXY, ~10% ASD) (Jha, Sheth, & Ghaziuddin, 2007; van Rijn, Bierman, Bruining, & Swaab, 2012), and 47,XYY syndrome (~20% ASD), but no increased rate in X chromosome trisomy (Bishop et al., 2011; Ross et al., 2012). In addition to the general association of aneuploidy with ID, these observations suggest several intriguing possibilities: 1) the Y chromosome is a risk factor for ASD, and 2) a second X chromosome is protective, possibly via genes that escape X-inactivation. Interestingly, the reported TS cases with co-morbid ASD

predominantly carry an intact maternal X chromosome, which led Skuse and colleagues to propose the theory that imprinted genes expressed only from the paternal X protect against ASD (Skuse, 2000). Since 40-50% of KS cases arise from maternal nondisjunction of the X, this subset of cases also lacks putative protective paternally expressed X genes, as do all 47,XYY cases. In combination with the presence of a Y chromosome, this lack of paternal X expression could then raise ASD risk for both syndromes, and for a higher proportion of XYY cases as is observed. However, larger epidemiological studies are needed to more accurately establish ASD prevalence in aneuploid individuals, and the parental origin of all sex chromosomes and patterns of escape from X-inactivation must be determined to better assess the validity of this model.

1.5: Sex hormonal contributors to ASD risk

One major theory that invokes a broad role for testosterone in ASD etiology is the Extreme Male Brain theory, which proposes that ASD arises from hypermasculinization of the brain (Baron-Cohen, 2002). A theory born from cognitive-behavioral observations, this masculinization is conceptualized along two cognitive dimensions: 1) empathizing, the drive to perceive others' feelings and thoughts and respond appropriately, and 2) systemizing, the drive to interact with and understand rule-based systems. Early work convincingly demonstrated that typical females score significantly higher on measures of empathizing and value placed on meaningful relationships with others (Baron-Cohen & Wheelwright, 2003, 2004), whereas typical males score significantly higher on measures of systemizing (Baron-Cohen, Richler, Bisarya, Gurunathan, & Wheelwright, 2003). In these studies, high-functioning ASD cases scored lower than typical males on measures of empathy and friendship, and higher than typical males on measures of systemizing.

Given that testosterone secreted by fetal testes during gestation drives human sexual differentiation to the male phenotype, Baron-Cohen and colleagues have proposed that fetal testosterone levels may also drive cognitive hypermasculinization in ASD. Findings of significant positive correlations between fetal testosterone levels and measures of systemizing (Auyeung et al., 2006) and autistic traits (Auyeung et al., 2009), and negative correlations with measures of empathizing (Chapman et al., 2006) and the quality of social relationships (Knickmeyer, Baron-Cohen, Raggatt, & Taylor, 2005) are consistent with this hypothesis. Recent work has even found a correlation between increasing fetal testosterone and volume of sexually dimorphic brain regions, specifically increased volume of the right temporoparietal junction/posterior superior temporal sulcus and decreased volume of the planum temporale/parietal operculum and posterior lateral orbitofrontal cortex volume (Lombardo et al., 2012). These results suggest that increased fetal testosterone levels predispose the differentiating brain to a hyper-masculine cognitive and neuroanatomical phenotype. Most recently, a large-scale study combined diagnostic history information from the Danish Historic Birth Cohort and Danish Psychiatric Registry with measurements of sex steroid hormone levels in banked amniotic fluid to show that increased levels of testosterone, progesterone, 17α-hydroxy-progesterone, and androstenedione at mid-gestation in males are associated with increased risk for ASD later in life (Baron-Cohen et al., 2014). This is the first evidence of a direct link between fetal testosterone levels and ASD risk, though the molecular and cellular mechanisms by which an elevated testosterone exposure can the development of an autistic brain remain unknown.

Interestingly, work from other investigators have suggested that testosterone beyond fetal development may also play a role in ASD pathophysiology. For example, levels of testosterone

and its precursors were found to be significantly elevated in a sample of ASD cases, with 57 of the 70 subjects having at least one androgen metabolite measuring above the upper limit of sex- and age-matched reference ranges (Geier & Geier, 2007). Subsequent studies have found increased androstenedione in serum from adults with ASD compared to controls irrespective of sex (Ruta, Ingudomnukul, Taylor, Chakrabarti, & Baron-Cohen, 2011), and a higher free androgen index in females with Asperger's syndrome versus controls (Schwarz et al., 2011), although a study of unaffected Japanese adults found no correlation between salivary testosterone levels and autistic-like traits (Takagishi et al., 2010). These findings were recently reviewed in detail by Geier and colleagues, collectively suggesting that hyperandrogenism may be a significant risk factor for ASD (Geier, Kern, King, Sykes, & Geier, 2012). They propose that more frequent assessment of testosterone levels in ASD cases is warranted to determine how prevalent this risk factor may be. Furthermore, testosterone levels or more general steroid dysfunction may constitute risk factors that override the typical protective mechanisms operating in females, as recent assessment of a large sample of females with ASD has found elevated rates of symptoms consistent with steroidopathic conditions as compared with controls (Pohl, Cassidy, Auyeung, & Baron-Cohen, 2014). Alternatively, these results may also suggest that it is the relative levels of steroid hormones, for example, the ratio between circulating estrogen and androgen levels, that are fundamentally responsible for males' increased risk for ASDs and for females' protection.

One potential pathway by which testosterone influences ASD risk may involve RORA (retinoic acid-related orphan receptor-alpha), a gene down-regulated in ASD lymphoblastoid cell lines (Nguyen, Rauch, Pfeifer, & Hu, 2010). RORA regulates expression of aromatase, the enzyme that converts androgens to estrogens, and is reciprocally activated by estradiol and

inhibited by testosterone (Sarachana, Xu, Wu, & Hu, 2011). These regulatory relationships may create a feedback loop that further elevates testosterone levels. Molecular pathways involving RORA may have more specific effects on brain as well, since RORA has a role in cerebellar and Purkinje cell development, and neuroprotection from oxidative stress. Another potential mechanism may involve immune system functioning in the brain, as a co-expression module of genes involved in immune system and glial function was observed to be up-regulated in adult autistic cortex (Voineagu et al., 2011), and sex hormones, particularly estradiol, have been shown to affect glial-neuronal interactions (McCarthy, Todd, & Amateau, 2003; Schwarz & Bilbo, 2012). Thus, it may not be the absolute levels of androgens or estrogens, but the balance between them that influences ASD risk.

Sex hormones are attractive candidates for sex-biased ASD risk and protective factors in that they raise the possibility for the development of treatments that cut across individuals' specific genetic liability. However, much work remains to determine the precise cellular and molecular mechanisms by which testosterone interacts with neurodevelopmental pathways and genetic risk loci to increase liability for autistic behavior, so that future treatments may specifically target these interactions. For example, in addition to the liability conferred directly to neural development by sex chromosomal genes, it also should be noted that sex chromosomal abnormalities frequently affect gonadal function. In fact, gonadal dysgenesis is common in TS and KS, causing abnormally low postnatal estrogen and testosterone levels, respectively, whereas testosterone levels are normal prenatally in KS and throughout life in 47,XYY cases (Price & van der Molen, 1970; Ratcliffe et al., 1994). While this more male-like hormonal environment may contribute to increased ASD risk in TS, hypogonadism in XXY males in the face of a nearly

10-fold increase in ASD prevalence suggests that the role of the hormonal milieu in ASD liability is likely complex and may be mediated by other risk factors.

1.6: Conclusions

ASD prevalence remains highly biased toward males, although more recent population screens have identified a higher proportion of autistic females relative to males than past work on clinically ascertained samples. Research is underway to better characterize the ways in which females present the autistic phenotype as compared with males, and further work is needed to determine if currently undiagnosed females would benefit from standard ASD services, as well as if diagnostic criteria will need to be adjusted to effectively identify these girls. ASD risk is likely to be multifactorial, with many different genetic variants and environmental factors contributing to liability, and still other sex-differential genetic and hormonal factors acting to potentiate risk to males and/or attenuate risk to females (Figure 1.1B). Evidence suggests that sex chromosomal gene dosage and sex hormone levels may be involved in setting sex-specific liability thresholds, but much future work is needed to definitively identify the most critical players at hand and to elucidate the precise mechanisms by which these sex-specific factors modulate presentation of the ASD phenotype.

In the following chapters, I describe two main approaches taken toward the identification of such ASD-modulatory sex-differential factors and mechanisms. First, I aimed to identify ASD risk variants with sex-differential penetrance or transmission patterns to serve as targets for future study of molecular mechanisms influencing sex-differential risk. For this aim, I have characterized sibling recurrence risk in multiplex families and twin concordance rates by sex in families from the Autism Genetics Resource Exchange (AGRE) to determine if risk in this

cohort follows a sex-differential threshold liability model. I have also performed a sex-stratified linkage analysis in subsets of these AGRE families to locate chromosomal regions potentially harboring sex-differential risk variants in each set of families, and I describe strategies for identifying specific risk variants within the implicated regions. Second, I aimed to characterize relationships between currently known ASD risk genes and sexually dimorphic processes in the typical human brain in order to begin to understand where, when, and how sex-differential biology modulates the impact of ASD risk variants. For this aim, I have evaluated sexually dimorphic gene expression patterns in post-mortem human brain from non-autistic subjects for overlap with the expression of known ASD risk genes, ASD-associated molecular pathways, and neural cell type markers. Together, these approaches investigate the interface between genetics and sexually dimorphic biology to advance our understanding of the roles that sex plays in the genetic architecture and general etiology of ASDs.

CHAPTER 2:

Recurrence rates provide evidence for sex-differential, familial genetic liability for autism spectrum disorders in multiplex families and twins

2.1: Abstract

Autism spectrum disorders (ASDs) are more prevalent in males, suggesting that females may be protected by a sex-differential mechanism and that autistic females may carry more, or more deleterious, genetic variants than affected males. Evidence from quantitative measures in population samples and at the genetic level in sporadic cases support this hypothesis, but reports on recurrence rates for ASD diagnoses have failed to observe significantly greater recurrence in siblings of earlier-born affected sisters. Here, I characterize recurrence patterns in AGRE (Autism Genetics Resource Exchange), a frequently studied ASD family collection including a large number of multiplex families, to determine if risk in these families also follows a female protective model. To identify sex differences in ASD risk and risk variant loads carried by affected children, I assess recurrence rates and associated quantitative traits in full siblings from multiplex nuclear families, as well as concordance in monozygotic and dizygotic twins, from AGRE. I compare recurrence rates and phenotypes between males and females, and between families or twin pairs with at least one affected female (female-containing, FC) and those with only affected males (male-only, MO).

I observe significantly higher ASD recurrence in male than female siblings, as well as significantly higher recurrence and concordance in FC than MO families and twin pairs, consistent with predictions from the female protective model. Males within FC families show a trend toward higher risk compared with FC females and MO males, but these differences do not reach significance. I also find significantly lower adaptive behavior scores in MO cases and a greater representation of females among higher functioning cases than expected, patterns that may be typical of multiplex families or specific to AGRE.

I conclude that by using families' full pedigree information to classify families *post hoc* as MO or FC, I am able to find recurrence rate differences between families carrying sex-differential familial risk loads that are consistent with the female protective model. However, when males and females are considered separately within MO and FC families, recurrence rates these subgroups approach 50%, consistent with a dominant inheritance pattern. This suggests a modification to the female protective model is necessary in which high-risk (here, multiplex) families can be sub-classified to those with and without effective female protective mechanisms, or alternatively where familial risk variants can be classified as penetrant or negligible in females. Under either model, future family-based genetic studies in AGRE may increase power or specificity by prioritizing families with affected females.

2.2: Background

Autism spectrum disorders are developmental disorders that appear early in life and are defined by impairments in social skills and language abilities, as well as restricted interests and repetitive behaviors (American Psychiatric Association, 2013). These symptoms present heterogeneously, with some autistic children showing severe intellectual disability and poor basic daily living skills, and others with high intelligence and capacity for independence. Overall, current prevalence estimates for ASD are approaching 1.5%, at 1 in 68 children, in the United States (Developmental Disabilities Monitoring Network, 2014), an increase that is thought to be attributed to increased awareness among parents, physicians, and teachers, leading to more frequent diagnoses. ASD diagnoses are approximately four times more frequent in males than females, and the mechanisms responsible for this sex difference are not well understood.

Genetic variation contributes strongly to ASD etiology, as evidenced by high concordance rates between twins (Bailey et al., 1995; Hallmayer et al., 2011) and high recurrence in siblings (Constantino et al., 2010; Ozonoff et al., 2011) as compared with the general population. ASDs also often appear sporadically as a result of deleterious *de novo* variants that arise in a parent's germ line. The identification of these rare, *de novo* copy number and single nucleotide variants (CNV, SNV) in ASD cases from simplex families has proven especially fruitful for risk gene discovery in recent years (Iossifov et al., 2012; Levy et al., 2011; Neale et al., 2012; O'Roak et al., 2012; Pinto et al., 2010; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007). However, the heritable or familial component of ASD's genetic risk architecture is still poorly understood, as family-based genetic linkage and association studies have identified very few replicable risk loci (Anney et al., 2010; Cantor et al., 2005; Liu et al., 2001; McCauley et al., 2005; Stone et al., 2004; Szatmari et al., 2007; Wang et al., 2009; Weiss et al., 2009; Werling, Lowe, Luo, Cantor, & Geschwind, 2014; Yonan et al., 2003).

It has been proposed that families with autistic children fall into two risk classes for ASD: a majority of low-risk families in which *de novo* variants are the primary genetic cause of ASD, and a minority of high-risk families in which inherited variants follow a dominant transmission pattern for males, with reduced penetrance in females (Ronemus, Iossifov, Levy, & Wigler, 2014; Zhao et al., 2007). Sibling recurrence risk estimates from multiplex families and from an infant sibling study fit this model, finding ASD recurrence in close to 50% of later-born male children in these families (Ozonoff et al., 2011; Zhao et al., 2007). These studies report far lower risk to later-born daughters from multiplex families of about 20%, consistent with the effects of a female-specific mechanism that protects females from manifesting an ASD phenotype.

This has been termed the female protective model, a variation on the multiple threshold liability model (Reich et al., 1975) for ASD risk in which genetic liability is distributed in the population and males and females have different thresholds, or minimum variant loads, at which they present an ASD phenotype (Tsai, Stewart, & August, 1981; Werling & Geschwind, 2013). Under this model, one would expect affected females to carry a greater risk variant load than affected males on average, and that this variant load, if inherited, would be shared among siblings. Recent evidence of higher scores on a quantitative measure of autistic traits in siblings of female probands as compared with siblings of male probands in two large, population-based samples supports this hypothesis (Robinson et al., 2013). Earlier work also observed trends toward higher rates of deleterious CNVs and SNVs among autistic females (Iossifov et al., 2012; Levy et al., 2011; Neale et al., 2012; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007), and a recent study found a significantly higher rate of these risk variants in females with ASD and other neurodevelopmental disorders, as well as preferential transmission of these CNVs from mothers (Jacquemont et al., 2014). However, these genetic studies all focused on sporadic ASD cases, and it is not known if the effects of increased variant loads in females and their siblings are also evident in multiplex families, particularly those from the Autism Genetics Resource Exchange (AGRE) collection.

AGRE is a collection of pedigree and phenotypic data and genetic material from families with autistic children (Lajonchere & AGRE Consortium, 2010). Due to an early focus on recruitment of families with multiple affected children, AGRE families have been widely utilized in genetic studies aiming to identify shared, familial risk variants, including linkage analyses (Cantor et al., 2005; McCauley et al., 2005; Stone et al., 2004; Szatmari et al., 2007; Weiss et al., 2009; Werling et al., 2014; Yonan et al., 2003) and family-based association testing (Anney et

al., 2010; Wang et al., 2009), and future work is likely to apply high throughput sequencing. Prioritizing families with affected females for this variant discovery work may enrich these samples for more deleterious and detectable variants with larger effect sizes than the average familial risk variant load. If ASD risk in AGRE families follows the female protective model, taking a female-focused approach for genetic studies may therefore facilitate discovery of key familial risk loci and advance our understanding of the combinations of, and interactions between, variants required to exceed the ASD liability threshold.

In order to establish whether AGRE families show transmission patterns consistent with the female protective model, here I have analyzed recurrence and concordance patterns in multiplex families and twin pairs. I aimed to determine whether ASD risk to twins and to siblings differs by the sex of the child being evaluated, as well as whether risk is greater in families with affected females. To assess this potential risk difference between families with different degrees of familial genetic liability, I classified families as female-containing (FC), with at least one affected female child, or as male-only (MO), with exclusively male affected children, by assessing each nuclear family's full pedigree. As compared with prospective studies for sibling recurrence, this approach allowed for a cleaner stratification of those families carrying risk loads penetrant only in males from families carrying putatively greater, female-penetrant risk loads.

I find that the recurrence patterns I observe are consistent with the concept of a female protective model in AGRE: recurrence risk is approximately two-fold higher in males compared with females, as well as two-fold higher in siblings from FC families compared with MO families. These relative differences in recurrence risk by sex and by family classification are also consistent across multiple subsets of families with different pedigree structures, suggesting that a

female protective model is evident in AGRE. Future studies focusing on families with affected females may therefore facilitate the discovery of loci and variants that contribute to the familial component of ASD risk architecture.

2.3: Materials and methods

2.3.a: Subjects

AGRE is a collection of phenotypic and genetic data from families with autistic children that was established in 1997 (Geschwind et al., 2001; Lajonchere & AGRE Consortium, 2010). Initially founded as a multiplex cohort, AGRE currently also includes simplex families, though it remains a valuable source of multiplex ASD families for study. Subjects in AGRE provided written informed consent or assent with parental agreement for behavioral evaluation and blood sample collection. This study was approved by the Western Institutional Review Board (AGRE), the Institutional Review Board at Washington University (subject recruitment, principal investigator: John Constantino), and by the Medical Institutional Review Board 3 at the University of California, Los Angeles.

Here, I focus separately on multiplex nuclear families with full sibling children and on twin pairs from AGRE. Starting from the catalogue of all AGRE subjects (database queried on April 14, 2014), which included 12,260 individuals from 2,278 families, I filtered families for inclusion in these analyses. Extended families were first parsed to nuclear families, and in order to enrich this sample for cases with genetic risk factors as opposed to environmental complications, nuclear families that included a child with pre- or perinatal insults, or premature birth before 35 weeks, were removed. Families with twin pairs or multiples of unknown zygosity were also excluded.

For this study, I classified as affected all subjects with study diagnoses of autism, "broad-spectrum," or "not quite autism" based on a clinician's evaluation of Autism Diagnostic Interview-Revised and Autism Diagnosis Observation Schedule scores. A "broad-spectrum" diagnosis is given to individuals with pervasive developmental disorders of varying severity, and includes subjects with conditions formerly termed PDD-NOS and Asperger's syndrome. A diagnosis of "not quite autism" is given to subjects who meet the autism cutoffs in all symptom domains but who do not meet the age of onset criterion, or conversely who meet the age of onset criterion but fall only one point short of autism cutoffs in one or more symptom domains. Families with one or more children with ambiguous diagnoses, in which AGRE clinicians did not evaluate a child but their parent reported a diagnosis from a community professional, were removed from analysis. All monozygotic (MZ) multiples and dizygotic (DZ) twin pairs from families meeting the above criteria that included at least one affected child were included in concordance analyses.

For analyses of siblings from multiplex families, I applied additional filters. First, since genetic risk variants carried by MZ multiples are non-independent, I selected one individual from MZ sets at random for inclusion. Families with only one affected child were then excluded, as were families in which affected children were half siblings. The birth order of all full sibling children was then assigned by sorting by the mother's or father's age at time of birth as available, or by individual subject ID for the 12 remaining families who lacked parental age information, as ID is typically assigned according to birth order. Each multiplex nuclear family was then classified by the sex of all of their affected children as either FC with at least one affected female child, or as MO with only affected male children. The final multiplex sample consisted of 5,328

individuals from 1,120 nuclear families, including 2,404 affected children, 684 unaffected full siblings, and 2,240 parents.

2.3.b: Sex ratios and recurrence risk

I calculated the ratio of males to females from all affected children in the multiplex family set. Then, since previous studies have shown differences in the relative numbers of affected males and females at different levels of functioning, I also calculated sex ratios within the subsets of affected children who met criteria for the strict autism diagnosis, children with lesser diagnoses of broad-spectrum or not quite autism, children with a Vineland Adaptive Behavior Scales (VABS) composite standard score within the top quartile in the sample (score ≥75), and children with a VABS score within the bottom quartile (score ≤50) (Sparrow, Balla, & Cicchetti, 1984).

To determine if multiplex families from AGRE show evidence of a female protective effect for ASD penetrance, I assessed ASD prevalence in siblings beyond the two affected children required per family to meet criteria for multiplex status. I then tested whether this risk differs by the sex of the evaluated children or by families' classification as MO or FC.

First, I assessed simple recurrence risk by recording the affection status of the first child born after the second affected child in each family with subsequent births (N=335 families), as well as the affection status of all children born after the second affected child (N= 453 children from 335 families). Next, since recurrence risk estimates may be influenced by parents' decisions to curtail their intended family size after having children with ASDs, referred to as stoppage, I also evaluated recurrence risk in families with specific structures. Following a 2007 study of recurrence risk in AGRE families (Zhao et al., 2007), I recorded the affection status of

the third child from families with exactly three children in which the first two children are affected with ASD (N=191 families). Here, I also expanded on this strategy to test the last born child in all families who had only one additional child after their second affected, regardless of total family size or birth order of the first two affected children (N=248 families).

To evaluate risk across multiplex families without limiting this analysis to later births in families who continued having children, and without weighting these estimates by including multiple children from large families, I next calculated what I refer to as "familial risk." In all families with at least three children, I calculated "familial risk" by running 100 trials in which two affected children are masked at random and affection status is evaluated in a third child. Familial risk is taken as the mean risk from these 100 randomizations. This method allowed for the inclusion of those families who stopped having children after their second affected, of all children in each family regardless of birth order, as well as ensuring that no family contributed disproportionately to the risk estimate since only one child per family was considered in each trial.

For each of these five estimates of recurrence or familial risk – (A) the next-born child from each family, (B) all subsequent children from all families, (C) the third-born child from three-child families, (D) the last-born child from families where the second affected child is born second to last, and (E) familial risk from 100 random selections of one child per family – I performed likelihood ratio tests in JMP (SAS Institute, Inc.) to compare differences in risk between males and females, and between MO and FC families. Since by definition there are no affected females in MO families, it was not possible to test for a significant interaction effect of sex by family classification. However, I did compare males' and females' risk within FC

families, and males' risk in FC to males' risk in MO families. P-values were adjusted for these 20 tests by Bonferroni correction.

I next extended the recurrence risk analyses to evaluate differences in the number of affected children per family between MO and FC families with at least three children, hypothesizing that families with an affected female would likely carry a greater, more penetrant genetic risk variant load and would therefore have a greater number of affected children. For comparison, I also used the proportion of males among all affected children in the final set of 1120 multiplex families (77.7%) to estimate the expected fraction of MO-classified families within families containing two, or three or more, affected children. Additionally, since previous work has reported increased ASD risk for children born shortly after affected siblings (Gunnes et al., 2013; Martin & Horriat, 2012), I evaluated the changes in recurrence between children born next and children born two places after the second affected child. As before, risk was evaluated separately by sex and family classification.

2.3.c: Quantitative phenotypes

Previous studies have reported an exacerbation of the male bias for ASD among high functioning individuals, and a greater representation of females among cases with intellectual disability (Banach et al., 2009; Fombonne, 1999; Volkmar et al., 1993; Yeargin-Allsopp et al., 2003). Therefore, in addition to the prevalence of ASD diagnoses, a binary trait, the severity of ASD and of intellectual and functional impairment may also differ by sex. Furthermore, if affected females do carry a greater genetic liability than males but are still protected from this liability to some degree, brothers of affected females may be more severely impacted than their sisters, and than brothers of affected males (MO families) (Robinson et al., 2013).

To address this, I tested several quantitative phenotypes related to ASD severity and intellectual ability, including the VABS composite standard score (Sparrow et al., 1984), the Peabody Picture Vocabulary Test (PPVT) standard score (Dunn, Bulheller, & Häcker, 1965), the Raven's Progressive Matrices estimated non-verbal intelligence quotient (Raven's NVIQ) (Raven & De Lemos, 1958), and the Social Responsiveness Scale (SRS) raw total score (Constantino, 2012), for sex differences overall and within FC families, and for differences between MO and FC families. I used scores as recorded by AGRE; in cases where a child was evaluated more than once, I used the most recent score for analysis. All scores recorded as "untestable" were set to missing. For the Raven's NVIQ, some children received scores of "ATN" (above the highest possible NVIQ score normalized by age, N=93) or "BTN" (below the lowest possible NVIQ score normalized by age, N=19). These scores were recoded as 160 and 20, which are above the observed maximum and below the observed minimum NVIQ scores in the remaining subjects. These high and low values match the maximum and minimum scores for the PPVT and VABS standard scores, two metrics that are scaled analogously to standard IQ.

Sex and family classification comparisons were assessed by t-tests allowing for unequal variances in JMP using the scores from one proband selected at random from each nuclear family. To test for sex-differential phenotypes within FC families, a paired t-test was used to compare scores from one randomly selected affected female and one affected male within each family. P-values were adjusted for 12 tests by Bonferroni correction.

2.3.d: Concordance in twin pairs

MZ multiples (111 twin pairs and 1 set of quadruplets) and DZ twins (193 pairs) with at least one affected member from families without perinatal complications or ambiguous

diagnoses were evaluated for ASD concordance. As for the multiplex families, each set of multiple births was classified as MO (male-male for MZ and DZ pairs, and male-female DZ pairs where only the male is affected) or FC (female-female for MZ and DZ pairs, and female-male DZ pairs where the female is affected), and concordance rates were compared between these groups using a likelihood ratio test in JMP.

2.4: Results

Evidence consistent with a female protective model for ASD has been reported at the population level (Robinson et al., 2013) and in sporadic cases from simplex families (Jacquemont et al., 2014). As ASD families from the AGRE cohort are widely used for genetic studies, particularly those targeting the familial components of genetic risk for ASD, here I aimed to determine whether patterns of ASD penetrance consistent with a female protective model are also evident in families from AGRE. Under a female protective model, or a higher genetic liability threshold for ASD in females than males, one expects to observe fewer affected females than males but higher recurrence within families that have at least one female with ASD. Thus, I assessed the ratio of affected males to females, recurrence risk patterns, and quantitative measurements of ASD severity and intellectual ability in multiplex nuclear families from AGRE. I also evaluated concordance patterns in MZ and DZ twin pairs from families in the AGRE collection.

Within 1,120 nuclear families with two or more full sibling children with study diagnoses of ASD, there are 2,404 affected children, including 1,867 affected males and 537 affected females for an overall male to female ratio in these families of 3.48:1 (Table 2.1). In contrast with reports of even greater male skew among less severely affected cases (Volkmar et al.,

1993), the sex ratio for children with lesser study diagnoses of broad-spectrum and not quite autism was 2.11:1, compared with 3.71:1 for children with study diagnoses of strict autism. A comparison of sex bias within cases scoring in the top and bottom quartiles from this sample on the Vineland Adaptive Behavior Scales (VABS), a measure of general functioning, showed a similar pattern, with a greater proportion of affected females falling in the high functioning quartile of the scale (M:F=2.35:1) than the lower functioning quartile (M:F=4.05:1). I note that these unexpected patterns may not accurately reflect trends at the general population level, and may instead be a consequence of the ascertainment scheme for AGRE. For example, parents who enroll their families in research studies may be more sensitive to the presentation of ASD symptoms in their female children than the average parent, thus increasing the rate of diagnoses among higher-functioning females in this sample.

Table 2.1: Ratio of affected males to females in multiplex families from AGRE

	No. families	No. affected children	No. affected males	No. affected females	Male:Female
All diagnoses	1120	2404	1867	537	3.48
Autism	1106	2158	1700	458	3.71
Spectrum	220	246	167	79	2.11
Top quartile (≥75) VABS	319	445	312	133	2.35
Bottom quartile (≤50) VABS	306	424	340	84	4.05

All diagnoses = autism, broad-spectrum, or not quite autism (NQA) study diagnoses. Spectrum = broad-spectrum or NQA. VABS = Vineland Adaptive Behavior Scales. VABS quartiles calculated from composite standard scores for affected children from AGRE multiplex families.

The recurrence rate for ASD in the multiplex set of 335 families with a child born after the second affected was found to be 39.4% (Table 2.2A). The recurrence rate in male children was 52.4% and 23.3% for females, a significant difference (P=7.64e-07) with a relative risk of 2.25 for males versus females (Figure 2.1A); these sex-differential rates closely match those observed by Zhao et al. (2007) in a subset of 165 AGRE families. I also found a comparable risk difference between FC and MO families, with 52.7% recurrence in FC and 28.9% recurrence in

MO families (P=1.80e-04) for a relative risk of 1.83 for FC families compared with MO (Figure 2.1B). Differences in recurrence risk between males and females from FC families, and between males from FC and MO families were not significant (Figure 2.1C). When all children born after the second affected child were included (N=453 children), as opposed to just one child per family, I observed similar patterns, with relative risk of 2.13 in males compared with females (P=5.80e-04) and 2.01 in FC compared with MO families (P=9.18e-04; Table 2.2B).

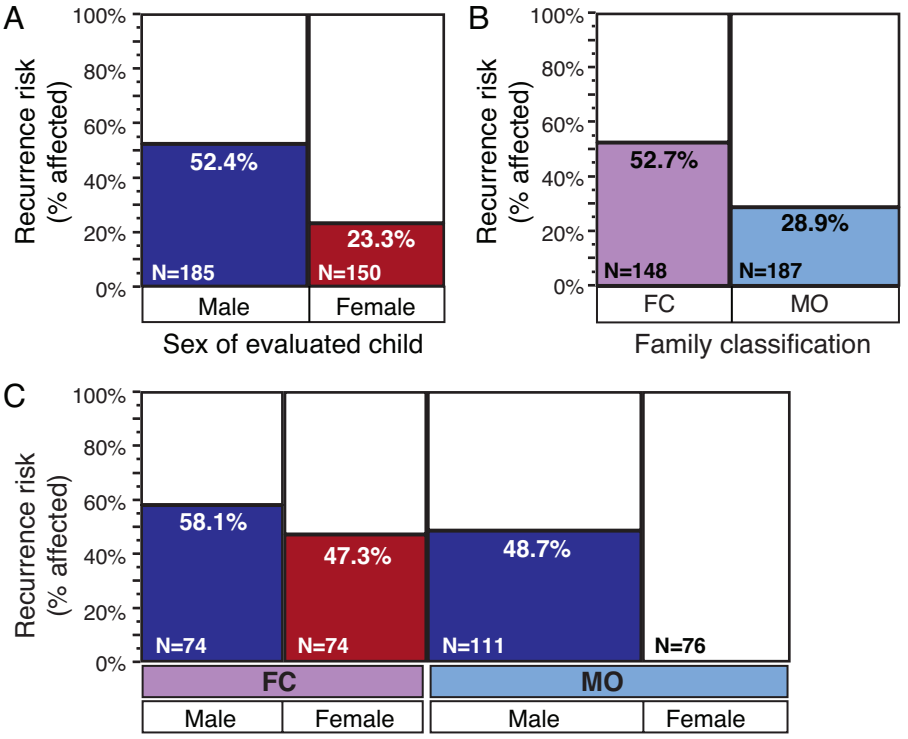


Figure 2.1: Recurrence risk for ASD in multiplex families from AGRE by sex and family classification. ASD risk in the first child born after the second affected child in each family (N=335 families) is highest for males and in female-containing families. Mosaic plots show the proportion of affected children by A) sex, B) family classification as FC (female-containing) or MO (male-only), and by C) sex within each family classification. Bar widths are proportional to the number of children in each group, which is also noted in the figure.

In any study of familial recurrence rates for a disruptive developmental condition such as ASDs, stoppage, or parents' decision to stop having children after an earlier born child is found to be affected, is frequently a concern (Hoffmann et al., 2014). Not only does stoppage limit the number of families with additional children from which to calculate recurrence risk, but it also may mask underlying differences in the families available for analysis. For example, autistic children in those families with multiple preceding or subsequent births may be less severely

Table 2.2: Recurrence risk in multiplex AGRE families by sex and family classification

A) First child born after 2nd affected. N=335 families, 39.4% affected

	Males	Females	Relative risk	Chisq	P-value	Adj. P-value
Males vs. Females	52.4%	23.3%	2.25	30.2	3.82e-08	7.64e-07
In FC	58.1%	47.3%	1.23	1.74	0.187	1
	FC families	**MO families**	**Relative risk**	**Chisq**	**P-value**	**Adj. P-value**
FC vs. MO	52.7%	28.9%	1.83	19.7	8.98e-06	1.80e-04
Males	58.1%	48.7%	1.19	1.60	0.206	1

B) All children born after 2nd affected. N=335 families, N=453 children, 36.2% affected

	Males	Females	Relative risk	Chisq	P-value	Adj. P-value
Males vs. Females	47.4%	22.3%	2.13	31.5	2.01e-08	5.80e-04
In FC	53.6%	44.6%	1.20	1.73	0.188	1
	FC families	**MO families**	**Relative risk**	**Chisq**	**P-value**	**Adj. P-value**
FC vs. MO	49.3%	24.6%	2.01	30.1	4.09e-08	9.18e-04
Males	53.6%	42.5%	1.26	3.08	0.0791	1

C) First child born after 2nd affected in families with exactly 3 children. N=191 families, 40.3% affected

	Males	Females	Relative risk	Chisq	P-value	Adj. P-value
Males vs. Females	53.3%	23.8%	2.24	17.5	2.90e-05	7.60e-05
In FC	65.8%	50.0%	1.32	2.00	0.157	1
	FC families	**MO families**	**Relative risk**	**Chisq**	**P-value**	**Adj. P-value**
FC vs. MO	57.7%	28.3%	2.04	16.6	4.59e-05	8.75e-05
Males	65.8%	46.4%	1.42	3.76	0.0525	1

D) First child born after 2nd affected is last child in family. N=248 families, 39.1% affected

	Males	Females	Relative risk	Chisq	P-value	Adj. P-value
Males vs. Females	51.8%	23.4%	2.21	21.4	3.80e-06	4.03e-07
In FC	61.2%	52.0%	1.18	0.859	0.354	1
	FC families	**MO families**	**Relative risk**	**Chisq**	**P-value**	**Adj. P-value**
FC vs. MO	56.6%	27.5%	2.06	21.1	4.37e-06	8.17e-07
Males	61.2%	46.6%	1.31	2.72	0.0993	1

E) Any 1 child selected after masking 2 affecteds at random, 100 randomizations (Familial risk). N=554 families, 17.3% affected

	Males	Females	Relative risk	Chisq	P-value	Adj. P-value
Males vs. Females	25.4%	9.8%	2.59	24.1	8.96E-07	1.79e-05
In FC	30.8%	22.2%	1.38	2.15	0.142	1
	FC families	**MO families**	**Relative risk**	**Chisq**	**P-value**	**Adj. P-value**
FC vs. MO	26.1%	11.4%	2.30	20.1	7.53E-06	1.51e-04
Males	30.8%	22.2%	1.39	2.47	0.116	1

FC = female-containing family; MO = male-only family. Chi-squared statistics and unadjusted P-values from likelihood ratio tests for affection status by sex or by family classification. Adjusted P-values have been corrected for 20 tests.

affected than autistic children in families of smaller size whose parents may have opted to limit their additional children. To address these potential underlying differences, I calculated

recurrence risk from 191 families with a specific structure: exactly three full sibling children, with affected first and second children, as was used to estimate recurrence rates in an earlier study of AGRE families (Zhao et al., 2007). Recurrence risk patterns in this specific set were comparable to those from all families with additional births, with 40.3% risk overall, a relative risk of 2.24 in males versus females (P=7.60e-05), and a relative risk of 2.04 in FC versus MO (P=8.75e-05; Table 2.2C). When this test is expanded and all families who had only one more child after their second affected are considered (N=248 families), recurrence risk again follows the same pattern, with relative risk of 2.21 in males versus females (P=4.03e-07), and relative risk of 2.06 in FC versus MO (P=8.17e-07; Table 2.2D). Differences in risk between males and females from FC families and for males from FC versus MO families were also observed, but none reach statistical significance.

In addition to recurrence risk, I also calculated familial risk by applying a randomization procedure that allowed me to include all 556 families with at least three full sibling children in the analysis. Familial risk in these families was found to be far less than recurrence risk, at just 17.5% (Table 2.2E). This is a likely consequence of including 221 families with unaffected, earlier-born children, though it may also be consistent with the presence of a maternal factor that acts to increase risk for ASD in children born after an affected child. The absolute familial risk estimates within each sex and family classification were similarly low, at 25.4% for males and 9.8% for females, and 11.4% for MO and 26.1% for FC families, but relative risks showed slightly more pronounced differences than for recurrence risk, at 2.59 for males versus females (P=1.79e-05), and 2.30 for FC versus MO (P=1.51e-04). Again, I observed differences between males and females within FC families and differences between males from MO and males from FC families, but these differences were not significant.

A logical consequence of higher recurrence risk in FC than MO families is a greater number of affected children in FC compared with MO families. In the 556 families with three or more full sibling children, families with only two affected children are more likely to be MO than FC (66.2% are MO versus 33.8% FC families), and families with three or more affected children are more likely to be FC than MO (62.1% are FC versus 37.9% MO families) (Figure 2.2). These fractions demonstrate a skew from expectation, as calculated by the probabilities that N affected children would all be males, using the estimate from this multiplex family set of 77.7% males among affected children. I observe a greater fraction of MO than expected among families with the minimum two affected children (60.3% expected, 66.2% observed), and fewer MO families than expected among those with three or more affected children (45.5% expected, 37.9% observed).

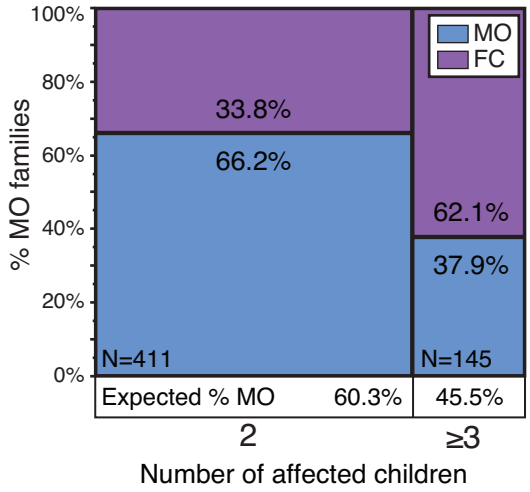


Figure 2.2: Number of affected children per family differs by family classification.
Families with 3 or more affected children are more likely to be classified as FC than MO, consistent with a higher recurrence risk in FC families. Mosaic plot shows the proportion of MO and FC families with at least 3 full sibling children that have only 2, or 3 or more, affected children. Bar widths are proportional to the number of families in each group, which is also noted in the figure.

Among full siblings who share a common source of genetic risk variation, I assume that risk for ASD is constant across the births in a family. However, recent work suggests that risk for ASD may be increased in later-born children (Martin & Horriat, 2012), perhaps due to advancing parental age (Kong et al., 2012) or another unknown maternal or paternal factor, or that risk for

ASD may be higher for children born shortly after an affected child (Gunnes et al., 2013). Therefore, I also evaluated recurrence risk in children born two births after the second affected child. I find no significant differences for risk to females and risk in FC families between the first and second births that follow a second affected child. However, I observe that risk to males and risk in MO families show significant decreases from the first to the second child born after the second affected child (males: X^2=6.72, p=0.0095; MO: X^2=8.90, p=0.0028; see Figure 2.3A-B). When children are stratified by both sex and family class, I find that risk to either males or females from FC families do not differ significantly between the first and second births after the second affected. In contrast, risk to males in MO families drops considerably for the second post-affected child, from 48.6% to 14.3% (X^2=9.51, p=0.0020; Figure 2.3C). These patterns are robust even when considering only the 87 families with at least two children born after their second affected child (Figure 2.3D-F).

To determine if males and females, and if MO and FC families, differed from one another in the presentation of ASD and its impact on functioning, I assessed quantitative measures of ASD severity, general functioning, and intellectual ability in the multiplex families. To compensate for the non-independence of multiple children from each family, I compared VABS, PPVT, Raven NVIQ, and SRS scores from one randomly selected proband per family. For within-FC family comparisons, I tested one randomly selected brother and sister from each family using a paired t-test. I observe trends toward lower VABS in males compared with females, higher SRS scores in MO compared with FC families, and higher Raven's NVIQ in FC males compared with their sisters, as well as significantly lower VABS scores in FC brothers compared with their sisters, though this difference is no longer significant after multiple testing correction (Table 2.3). I do find significantly lower VABS scores in probands from MO

compared with FC families (P=0.014), indicating that adaptive behavior in affected children from FC families is not as severely impacted, on average, as affected children from MO families.

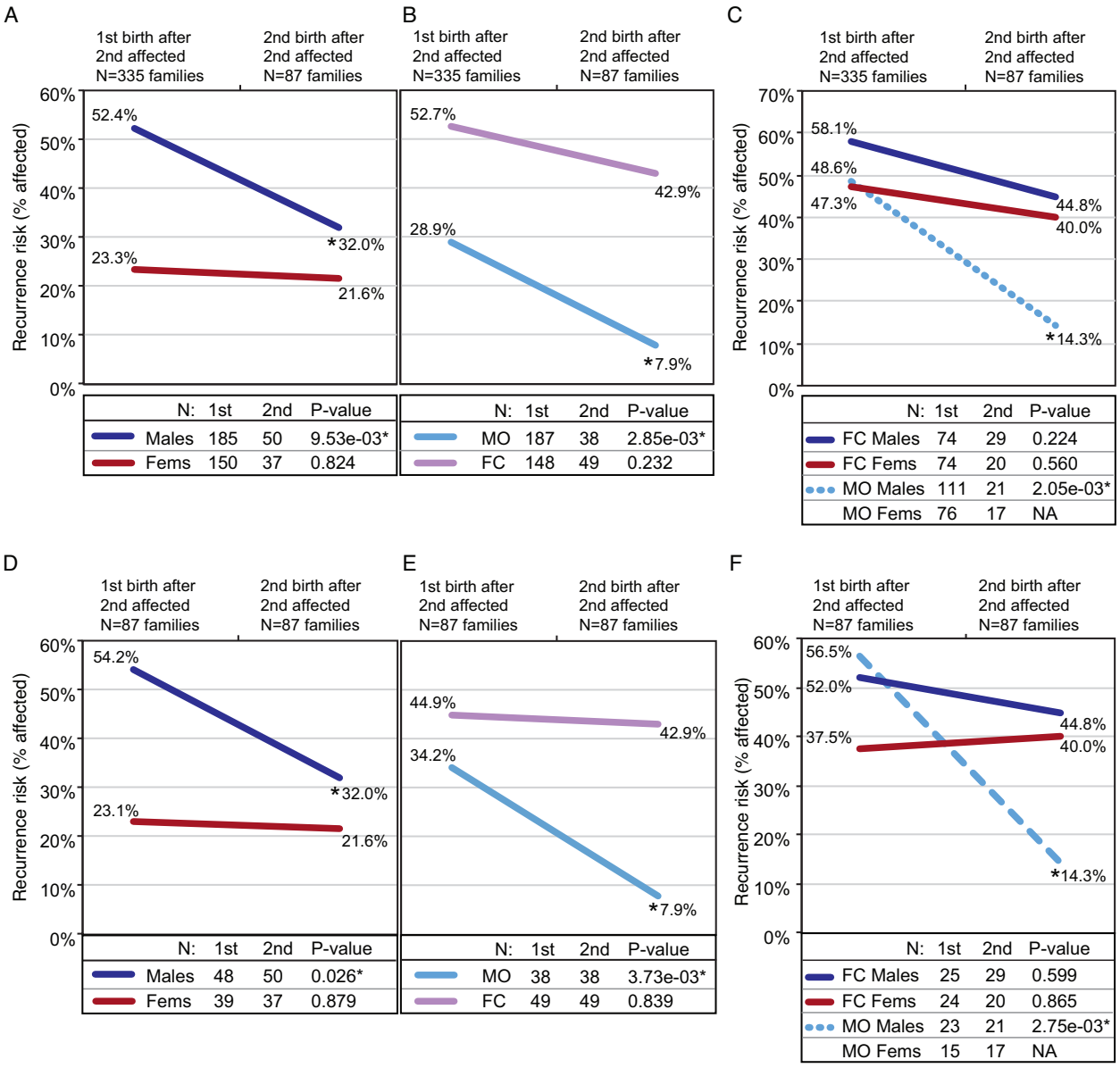


Figure 2.3: Patterns of ASD risk to children born after the multiplex criterion is met.
Line graphs show the proportion of affected children born first and second after a second affected child in all available families with subsequent births by A) sex and family classification as FC (female-containing) or MO (male-only), and by B) sex within each family classification. By definition, 0% of females from MO families are affected. C) and D) show proportions of affected children for only those families with at least two children born after their second affected child. The number of families evaluated in each group and the P-values from likelihood ratio tests comparing ASD recurrence rates in the first versus the second child born after the second affected is shown in tables below each figure panel.

Table 2.3: Adaptive behavior, intellectual ability, and ASD symptoms by sex and family classification

A) Males vs. females, 1 randomly selected proband per family

	Males			Females			Males vs. females				
	N	Mean	SD	N	Mean	SD	Diff.	SE	t	P	Adj. P
VABS	587	60.90	18.62	177	64.34	20.92	-3.43	1.75	-1.96	0.051	0.616
PPVT	505	85.34	26.68	153	85.73	25.23	-0.39	2.36	-0.17	0.868	1
Raven's NVIQ	464	104.27	23.04	145	101.09	22.20	3.18	2.13	1.49	0.138	1
SRS	373	103.85	33.34	120	101.53	37.33	2.32	3.82	0.61	0.544	1

B) MO vs. FC, 1 randomly selected proband per family

	MO			FC			MO vs. FC				
	N	Mean	SD	N	Mean	SD	Diff	SE	t	P	Adj. P
VABS	459	59.84	18.29	305	64.51	20.24	-4.67	1.44	-3.25	0.001	0.014
PPVT	394	84.20	27.38	264	87.28	24.63	-3.08	2.05	-1.50	0.133	1
Raven's NVIQ	357	103.85	24.13	252	103.03	20.98	0.81	1.84	0.44	0.658	1
SRS	293	105.64	32.98	200	99.83	36.02	5.81	3.19	1.82	0.070	0.834

C) Males vs. females within FC families, 1 randomly selected affected brother & sister, paired test

	Males			Females			Males vs. females				
	N	Mean	SD	N	Mean	SD	Diff	SE	t	P	Adj. P
VABS	264	62.73	18.65	264	66.03	20.12	-3.30	1.30	-2.55	0.012	0.138
PPVT	180	89.85	25.06	180	89.77	24.00	0.08	2.22	0.04	0.970	1
Raven's NVIQ	171	107.06	20.92	171	103.47	22.91	3.60	2.16	1.67	0.097	1
SRS	149	107.81	32.88	149	103.81	35.53	4.00	3.70	1.08	0.282	1

FC, female-containing family; MO, male-only family; VABS, Vineland Adaptive Behavior Scales composite standard score; PPVT, Peabody Picture Vocabulary Test standard score; Raven NVIQ, Raven's Progressive Matrices estimated non-verbal intelligence quotient; SRS, Social Responsiveness Scale raw total score. T statistics and unadjusted P-values from t-tests allowing for unequal variances for affection status by sex or by family classification. C) used a paired t-test, comparing affected males and females from the same families. Adjusted P-values have been corrected for 12 tests.

As an originally multiplex cohort, AGRE includes a large number of affected twin pairs and other multiple births. To determine if the female protective model is supported by twin concordance, I assessed concordance rates in MZ multiples and DZ twin pairs from AGRE. I identified high concordance rates in MZ multiples, with male pairs (MO) showing 95.7% concordance and female pairs (FC) showing 85.0% concordance; this sex difference in concordance rates was not significant (Table 2.4). For DZ twin pairs, I find a significantly higher concordance rate among FC pairs (63.3%) than among MO pairs (36.1%; P=9.31e-04), a pattern that is consistent with the recurrence risk estimates from non-twin siblings. I also note that in 56 of the 62 discordant male-female DZ twin pairs, it is the male who is affected. This degree of

male bias approximates the expected pattern under a female protective model, given a 4.3:1 M:F for ASD among DZ twins from AGRE (N expected for male-affected, discordant DZ pairs = 54 pairs).

Table 2.4: Concordance rates in twin pairs by sex of affected twins

	No. pairs	No. FC pairs	No. MO pairs	FC concordant	MO concordant	Relative risk	Chisq	P-value
MZ twins	112	20	92	85.0%	95.7%	0.89	2.60	0.110
DZ twins	193	49	144	63.3%	36.1%	1.75	11.0	9.31e-04

FC = female-containing family; MO = male-only family; MZ = monozygotic twins; DZ = dizygotic twins. Chi-squared statistics and unadjusted P-values from likelihood ratio tests for affection status by MO and FC classification.

2.5: Discussion

I have evaluated recurrence risk and concordance for ASDs in multiplex families and twin pairs from AGRE, one of the largest and most frequently studied ASD family collections available, for evidence of a female protective model. Under this model, females are less frequently affected with ASD than males, but those females who are autistic are expected to carry a greater genetic liability, or risk mutational load, than males. In multiplex families, where it is assumed that multiple affected family members inherit the same risk variants, it is expected that affected females will share this high liability with their siblings, increasing recurrence rates in these families as compared with siblings of affected males. In families from AGRE, I observe recurrence and concordance patterns consistent with this model, with approximately two-fold increased risk to males compared with females, as well as two-fold increased risk to siblings and co-twins from families or twin pairs with affected females. These findings of a female protective effect in multiplex families from the AGRE family collection can best inform future study design using this cohort for discovery of familial risk variants.

Across epidemiological studies in different countries, at different times, and using different diagnostic criteria, ASD has reliably shown a greater prevalence in males compared with females (Fombonne, 2009), suggesting that males are at increased risk for, or females are protected from, ASD. It has also been proposed that families fall into two risk classes: a low-risk class in which families' shared genetic background does not contain causal risk variants, and a high-risk class including multiplex families in which family members transmit risk variants in a dominant pattern (Ronemus et al., 2014; Zhao et al., 2007). Here, the observation of a ~52% recurrence rate in males is consistent with this notion of multiplex families belonging to the high-risk class with dominantly expressed risk variation, but the observation of a greater than two-fold reduction in females' recurrence to just ~23% also illustrates the presence a female-protective factor that operates even in these high-risk families, which is consistent with recurrence rates observed by Zhao et al. (2007). This significant difference in risk to males and females agrees with other studies of recurrence risk in infant siblings (Ozonoff et al., 2011) and in families from AGRE and the Interactive Autism Network registry (Constantino et al., 2010).

To examine the validity of a multiple threshold liability model and related female protective effect in multiplex AGRE families, I also tested for differences in recurrence and concordance rates between families with only affected males, hypothesized to carry relatively lesser genetic liability for ASD, and families with at least one affected female, hypothesized to carry relatively greater genetic liability for ASD. Under these hypotheses, one expects to find higher recurrence in children from FC families than from MO families, and this is what I observe: when male and female siblings are considered together, FC multiplex families show recurrence rates at or above 50% in all multiplex family groups examined, while recurrence rates in MO families fall between 24-28% (Table 2.2).

This difference in risk may be attributed to several potential scenarios. One possibility is that, in high-risk families, autistic female children's daily functioning is not as severely impacted as males'. This comparatively reduced parental burden across FC families may be associated with less stoppage and an increased number of children born in FC families who are at high risk for ASD. Parents in the MO families who continued having children after their second affected child may have different motivations for doing so, and since recurrence risks are calculated from later-born children, differences in their estimates may be affected by these inherent group differences.

Of the four quantitative phenotypes tested for group differences, I observe significantly lower VABS scores in affected children from MO families as compared with FC families (difference=-4.67 points, standard error=1.44, adj. P=0.014). I also observe a marginally significant increase in the proportion of FC families who continued having children after their second affected (likelihood ratio test chisq=4.25, P=0.039), though this difference is less statistically significant when only families with three or more total children are considered (chisq=4.28, P=0.077). Together, these trends suggest that the adaptive behavior of affected children from FC families may be less severely impacted by the risk variants responsible for their ASD, and that their parents are more likely to continue having children despite having multiple earlier-born children on the spectrum. A comparison of quantitative phenotypes for children from families who stopped versus continued having children after their second affected child does show a trend for lower VABS scores in families who stopped, but this difference is not significant for all families or for FC families after multiple testing correction (Table 2.5).

Another possible explanation for the two-fold higher risk to siblings in FC families is that predicted directly by the multiple threshold liability model, wherein affected females carry, and

Table 2.5: Adaptive behavior, intellectual ability, and ASD symptoms by family stoppage status

A) Families who stopped vs. continued having children, 1 randomly selected proband per family

	Stop			Continue			Stop vs. continue				
	N	Mean	SD	N	Mean	SD	Diff.	SE	t	P	Adj. P
VABS	537	60.9	19.0	227	63.5	19.6	-2.59	1.54	-1.68	0.094	0.749
PPVT	462	84.4	26.9	196	87.9	24.9	-3.56	2.18	-1.64	0.102	0.818
Raven NVIQ	422	103.2	23.6	187	104.2	21.1	-0.96	1.93	-0.50	0.620	1
SRS	356	101.1	33.7	137	109.0	35.4	-7.90	3.51	-2.25	0.026	0.204

B) Families with ≥3 full sibling children who stopped vs. continued having children, 1 randomly selected proband per family

	Stop			Continue			Stop vs. continue				
	N	Mean	SD	N	Mean	SD	Diff.	SE	t	P-value	Adj. P
VABS	142	58.1	20.2	227	63.5	19.6	-5.40	2.13	-2.53	0.012	0.096
PPVT	115	82.4	30.3	196	87.9	24.9	-5.52	3.34	-1.65	0.100	0.799
Raven NVIQ	113	101.4	27.1	187	104.2	21.1	-2.78	2.98	-0.93	0.352	1
SRS	95	102.2	35.1	137	109.0	35.4	-6.74	4.70	-1.43	0.153	1

C) FC Families with ≥3 full sibling children who stopped vs. continued having children, 1 randomly selected proband per family

	Stop			Continue			Stop vs. continue				
	N	Mean	SD	N	Mean	SD	Diff.	SE	t	P	Adj. P
VABS	45	56.5	22.8	109	67.0	19.7	-10.42	3.89	-2.68	0.009	0.109
PPVT	35	86.1	28.1	96	89.6	23.7	-3.54	5.34	-0.66	0.510	1
Raven NVIQ	38	101.3	23.9	91	103.8	21.1	-2.51	4.46	-0.56	0.575	1
SRS	33	107.8	30.4	57	102.8	37.6	4.92	7.26	0.68	0.501	1

VABS = Vineland Adaptive Behavior Scales composite standard score; PPVT = Peabody Picture Vocabulary Test standard score; Raven NVIQ = Raven's Progressive Matrices estimated non-verbal intelligence quotient; SRS = Social Responsiveness Scale raw total score. T statistics and unadjusted P-values from t-tests allowing for unequal variances for family's stoppage status. Adjusted P-values are corrected for 12 tests.

share with their siblings, a genetic risk load of greater magnitude than is carried by the average affected male. If it is the case that affected females in multiplex families carry a quantitatively more deleterious liability load than males, one would expect to observe higher recurrence in brothers from highly loaded FC families than from lesser-loaded MO families. While I do observe a 1.19- to 1.42-fold higher risk to males in FC families across different family sets (Table 2.2), this increase is not statistically significant. Additionally, I would expect to observe more frequent recurrence or more severe phenotypes in males than females from the highly loaded FC families, since by definition males lack female-specific protective mechanisms. I do find recurrence rates and phenotypic differences consistent with this expectation, with a 1.18- to

1.32-fold increase in ASD risk to FC males compared with females and a 3.3-point average decrease in males' VABS score compared with their affected sisters (Table 2.3). Again, neither difference is statistically significant. Furthermore, one would expect unaffected males and females to be more frequently observed in MO than FC families, since MO families are predicted to carry lesser heritable genetic risk loads than FC families. While I observe a greater than expected proportion of MO families with two or more unaffected female children (total 59 families, 71.2% MO observed vs. ~60% expected), families with two or more unaffected male children show no such skew (total 46 families, 60.9% MO observed vs. ~60% expected).

Instead, sex-specific recurrence rates within each family class suggest two potential modifications to the multiple threshold liability model. The first possibility is that high-risk families fall into two subclasses: 1) families in which female-protective factors are robust and only male children are affected, and 2) families in which female-protective factors are absent or attenuated, leaving females as vulnerable as their brothers. The second possibility is that females are protected from the effects of only a subset of specific inherited risk variants, and are as vulnerable as males to the effects of others. In other words, it may also be that the heritable genetic risk variants that push families into the high-ASD-risk class can be assigned to one of two classes: 1) variants that are penetrant in both males and females, and 2) variants that are predominantly penetrant in males.

While I observe (non-significant) shifts in recurrence risk consistent with sex-differential genetic risk loads, recurrence rates in males from both MO and FC families, and in females from FC families, are very close to 50%. This is the recurrence rate expected under a dominant inheritance model, which fits the genetic architecture proposed by Zhao and colleagues for multiplex families (2007). Thus, these results are consistent with the second proposed model

above, where familial risk variants with dominant effects belong to either a sex-neutral or sex-biased class. This is because (though dependent on the precise mechanisms involved) female protective factors may be less likely to be uniformly impacted across all children in a family, and this non-uniform impact would likely skew transmission rates away from the 50% expectation under a purely dominant model. However, with these data I cannot rule out the family sub-classification model under which female-protective factors are absent or attenuated in the children of all, or a subset of, FC families.

It is also interesting to note that risk to later-born children shows different patterns in MO and FC families. FC families show fairly constant recurrence rates across the two births following the second affected child for both male and female children, but MO families show a large drop in risk to the second child born after the second affected child. Though the number of families with at least two children born after their second affected child is small, this pattern suggests a potential contribution from transient, non-heritable factors, perhaps related to the maternal uterine environment, that predominantly impact males.

Discordance patterns in DZ twins are also consistent with this concept of non-heritable risk factors with male-biased effects, as female DZ twins are far less likely to be affected with ASD than their male co-twins. In fact, twinning itself may be a risk factor for ASD (Betancur, Leboyer, & Gillberg, 2002; Greenberg, Hodge, Sowinski, & Nicoll, 2001), and the risk associated with twinning may differ by sex. For example, studies of neonatal outcomes show that male twins are preferentially vulnerable to respiratory distress (Steen, Kallen, Marsal, Norman, & Hellstrom-Westas, 2014), and that males born preterm or with low birth weight, conditions that are more common among twins than singleton births, require more respiratory, circulatory,

and pharmacological support than females (Stevenson et al., 2000; Elsmen, Pupp, & Hellstrom-Westas, 2004).

In contrast with the results from this study, previous investigations of recurrence risk have tested for and failed to observe a significant effect of the sex of older affected siblings on risk to later-born children (Constantino et al., 2010; Gronborg, Schendel, & Parner, 2013; Ozonoff et al., 2011; Sandin et al., 2014). However, in contrast with these reports, I do not apply a prospective approach for my recurrence risk calculations, and instead classify families based on the sex of all of their affected children regardless of birth order. By considering each family's completed pedigree, I am able to achieve a cleaner delineation between families with female-penetrant and male-specific risk loads. With a prospective approach, a greater number of families carrying female-penetrant variants but with early-born affected males would be misclassified as MO, thus obscuring recurrence risk differences between the two family classes. While the post-hoc classification approach used here is not directly translatable to clinical settings and family planning decisions, it is highly informative for genetic study design, since the birth of an affected female child at any time in a family's pedigree is a positive indicator of a high familial liability load that is likely to have a larger effect size than heritable variants in MO families on average. Genetic studies that focus on FC families may therefore have increased power to detect heritable risk variants, which have so far remained largely elusive.

Previous studies also report recurrence risks of 10-20% overall, far lower than is seen here. These lower frequencies can be attributed to the joint consideration of families with both multiplex- and simplex-typical genetic architecture. Interestingly, these mixed recurrence rates approximate my overall familial risk estimate of 17.3% ASD prevalence among all siblings of affected children. Therefore, it may be that there is also a mix of genetic risk architectures among

multiplex families from AGRE that is masked when only later-born children are considered. Or more likely, this lower familial risk estimate is a consequence of ascertainment, since it adds to the risk calculation 221 families with an unaffected child born before, or between, two affecteds who are not used when estimating recurrence in multiplex families. Including these unaffected children therefore reduces the risk estimates, though the relative risks between males and females, and FC and MO families, are the same or greater for familial than recurrence risk, likely due to the consideration of siblings that are comprised of a greater proportion of unaffected females than males (Table 2.2E).

As a family cohort that is frequently utilized for family-based genetic studies, it is important to note the ways in which AGRE families may differ from the general population of families with autistic children. One such difference is in the average level of functioning of affected females. Previous work has reported closer to equal representation of autistic males and females among severely impacted cases with comorbid intellectual disability, and a more pronounced male bias among high-functioning individuals; in AGRE multiplex families, I observe the opposite pattern. Affected males are more strongly overrepresented among cases with low adaptive behavior and strict autism diagnoses, and affected females are more equally represented among cases with lesser, spectrum diagnoses and higher adaptive behavior (Table 2.1). The reasons for this difference are not known. It has been suggested that current diagnostic tools are calibrated to a male-typical phenotype and that females are under-diagnosed for ASD as a result of non-prototypical presentation of ASD symptoms in females. Parents with an affected child may likely be more aware of ASD symptoms than the average parent, and this increased awareness may lead to higher rates of diagnosis among their daughters than for sporadic female cases.

I also caution that the MZ and DZ twin concordance rates reported here are not likely reflective of the general population. This is because a large proportion of AGRE families were recruited for their multiplex status, an ascertainment strategy that would likely have lead to the preferential recruitment of families with concordant twin pairs. Finally, I note that the burden of care required by an autistic child can be substantial, and so in addition to potential differences at the genetic risk level, there may be key differences between those parents who continue to have children after an their earlier-born child is diagnosed with ASD, and parents who do not. Though they may carry highly penetrant, heritable risk variants, the latter families will appear as simplex and are therefore not characterized here.

2.6: Conclusions

To investigate sex differences in ASD risk and the implications of these sex differences for familial genetic risk architecture, I have characterized recurrence risk in multiplex families from the AGRE collection. I observe significant sex differences in recurrence rates, with females showing a two-fold reduction in risk compared with males' risk. I also observe two-fold higher recurrence rates in families with at least one affected female as compared with families whose affected members are exclusively male, a significant difference that is expected under the sex-differential threshold liability model for ASD. However, the approximately 50% recurrence rate in each class of affected children (males from MO families, females from FC families, and males from FC families) is more consistent with modified multiple threshold liability models, in which either 1) families are stratified by the relative presence or absence of female-protective factors, or 2) risk variants are classified as penetrant or impervious in females (or, a combination of both). Under either updated model, genetic study designs that stratify families by the sex of their

affected children or that prioritize families with affected females will likely be poised for discovery of heritable variants with larger effect sizes or sex-differential penetrance, both of which will be key for a full understanding of the shared familial component of ASD risk.

CHAPTER 3:

Identification of suggestive sex-differential risk loci and replication of linkage at chromosome 20p13 for autism spectrum disorder

3.1: Abstract

Autism spectrum disorders (ASDs) are male-biased and genetically heterogeneous. While sequencing of sporadic cases has identified *de novo* risk variants, the heritable genetic contribution to ASD risk and the mechanisms driving the male bias are less understood. Here, we aimed to identify inherited and sex-differential risk loci in the largest available, uniformly ascertained, densely genotyped sample of multiplex ASD families from the Autism Genetics Resource Exchange (AGRE), and to compare results to those from earlier studies of AGRE. From a total sample of 1008 multiplex families, we performed genome-wide, non-parametric linkage analysis in a discovery sample of 847 multiplex families and separately on subsets of families with only male affected children (male-only, MO) or with at least one affected female child (female-containing, FC). Loci showing evidence for suggestive linkage (logarithm of odds, LOD>2.2) in this discovery sample or in previous AGRE samples were re-evaluated in an extension study utilizing the total available multiplex sample of 1008 families. For regions with evidence of genome-wide significant linkage signal in the discovery stage, those families not included in the corresponding discovery sample were then evaluated for independent replication of linkage. Association testing of common single nucleotide polymorphisms (SNPs) was also performed within regions of suggestive linkage. We observed an independent replication of previously observed linkage at chromosome 20p13 ($p<0.01$), while loci at 6q27 and 8q13.2 showed suggestive linkage in our extended sample. Suggestive sex-differential linkage was observed at 1p31.3 (MO), 8p21.2 (FC), and 8p12 (FC) in our discovery sample, and the MO signal at 1p31.3 was supported in our expanded sample. No sex-differential signals met replication criteria, and no common SNPs were significantly associated with ASD within any identified linkage regions. With few exceptions, analyses of subsets of families from the AGRE

cohort identify different risk loci, consistent with extreme locus heterogeneity in ASD. Generally, large samples appear to yield more consistent results, and sex-stratified analyses facilitate the identification of sex-differential risk loci, suggesting that linkage analyses in large cohorts are useful for identifying heritable risk loci. Additional work is needed to identify the specific variants responsible for increasing ASD risk, including targeted re-sequencing and much larger samples.

3.2: Background

Autism spectrum disorders (ASDs) are a group of neurodevelopmental conditions characterized by severe social impairment that affect 1 in 68 individuals (Developmental Disabilities Monitoring Network, 2014). Genetic factors have long been known to contribute significantly to ASD risk based on twin studies (Hallmayer et al., 2011), the sibling recurrence risk (Constantino et al., 2010; Ozonoff et al., 2011), and elevated rates of comorbid ASD in populations with a wide variety of monogenic syndromes such as Fragile X or Timothy Syndrome (Abrahams & Geschwind, 2008; Berg & Geschwind, 2012). ASDs are also known to present heterogeneously across the population of affected individuals, and results from recent genetic studies strongly suggest that genetic risk factors for ASD are similarly diverse. Namely, copy number variant (CNV) and exome sequencing studies of sporadic ASD cases from single-incidence ("simplex") families have found numerous novel, *de novo* risk variants (Iossifov et al., 2012; Levy et al., 2011; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007), and no significant signal for rare inherited variation. Estimates based on these findings project that approximately 1000 genes are likely to contribute to ASD etiology.

While a highly productive approach for gene discovery, the study of simplex families is designed to identify mostly the non-inherited genetic component of ASD risk: rare variants resulting from *de novo* mutations, in which variants arise in the germ cell and are not carried by the mother or father. However, evidence of high heritability for ASD (Klei et al., 2012), high sibling recurrence risk (Constantino et al., 2010; Ozonoff et al., 2011), and aggregation of subthreshold ASD-like phenotypes in families (Bailey et al., 1995; Constantino & Todd, 2005; Folstein & Piven, 1991; Piven, Palmer, Jacobi, Childress, & Arndt, 1997) suggest that inherited genetic variation also plays a significant role in ASD etiology. Additionally, while germline mutations, potentially shared between affected siblings, may also plausibly affect ASD risk in multiplex families, current evidence suggests that rare *de novo* CNV events are more prevalent among sporadic cases than cases from multiple-incidence ("multiplex") families (Sebat et al., 2007). Largely however, the specifics of ASD's genetic architecture that differ between simplex and multiplex families are unknown. Therefore, studies of familial transmission to identify regions of genetic linkage in multiplex families remain an important approach to identifying predisposing genes.

Another important clue to ASD etiology lies in its consistently male-biased prevalence (Fombonne, 2009). There is an approximately 4:1 male bias, a phenomenon that is likely driven, or at least influenced, by the actions of sex-specific biological factors, such as sex chromosomes or steroid hormones that potentiate and attenuate ASD risk in males and females, respectively (Werling & Geschwind, 2013). Indeed, several ASD and intellectual disability risk genes have been identified on the X chromosome (Abrahams & Geschwind, 2008; Gecz, Shoubridge, & Corbett, 2009), including FMR1 (Bailey, Jr. et al., 1998), NLGN4X, and NLGN3 (Jamain et al., 2003), demonstrating that in some cases ASD may be X-linked. However, the proportion of ASD

cases currently attributable to X-linked variants remains insufficient to account for the degree of male bias observed in ASD prevalence. The notion of female protective factors on a broader scale is supported by the observation of an increased proportion of autistic females relative to males carrying variants of large effect size, such as large CNVs or deleterious single nucleotide variants (SNVs) (Levy et al., 2011; Sanders et al., 2011; Sebat et al., 2007). However, with few exceptions (Sato et al., 2012), it is unknown which specific autosomal risk variants are differentially penetrant by sex, thus contributing to the sex bias in ASD prevalence. For example, a greater number of variants may be associated with ASD risk in males as compared to the number of variants that also, or specifically, confer risk to females. The discovery of such sex-differential risk loci would provide genetic clues for investigation of the biological mechanisms driving the ASD male bias. However, because these signals are likely masked by heterogeneity within sex-mixed cohorts, stratification of multiplex ASD family cohorts by proband sex may facilitate the identification of novel, sex-differential loci harboring inherited risk variants.

Most previous linkage studies of ASD have used relatively small samples (<350 families) and markers with coarse resolution and incomplete information (Barrett et al., 1999; Cantor et al., 2005; Lamb et al., 2005; Liu et al., 2001; McCauley et al., 2005; Schellenberg et al., 2006; Stone et al., 2004), and there has been little agreement between studies in the reported findings. Furthermore, the larger studies (Szatmari et al., 2007; Weiss et al., 2009) combine subjects drawn from several diverse populations with different ascertainment schemes. Multiplex family samples ranging from 109 to 753 families from the AGRE cohort have been previously tested for linkage (Cantor et al., 2005; Liu et al., 2001; McCauley et al., 2005; Stone et al., 2004; Szatmari et al., 2007; Weiss et al., 2009; Yonan, et al., 2003), and the linked risk loci reported by these

analyses, with a few exceptions (Cantor et al., 2005; Stone et al., 2004), show little agreement. This may be due to genetic heterogeneity, small sample sizes, or sparse marker coverage.

In this study, we used a pruned set of single nucleotide polymorphism (SNP) markers providing nearly complete linkage information on all autosomes and the X chromosome (information content greater than 0.976 for 99.5% of regions covered by SNP genotypes) for our analyses in the largest available AGRE sample. Specifically, we performed non-parametric linkage testing in three stages to 1) identify novel risk loci in a discovery sample of 847 multiplex families, 2) to confirm loci identified in this discovery sample, or reported by earlier analyses of AGRE samples, in an extended sample of 1008 multiplex families, and 3) to test for formal replication of genome-wide significant linkage signals in the independent portion of families not tested in the corresponding discovery studies (Figure 3.1).

For all stages, we also applied a stratification approach used previously (Cantor et al., 2005; Stone et al., 2004; Szatmari et al., 2007) to identify sex-differential loci, now using a sample three times as large as the original and a panel of markers that are substantially more informative (Stone et al., 2004). We then used the family-based association transmission disequilibrium test (TDT) on genotyped and imputed SNPs within linked regions to identify common variants conferring increased risk for ASD. With the addition of 343 AGRE families beyond those tested in the most recent linkage analysis using subjects from AGRE (Weiss et al., 2009), this comprises the largest linkage study for ASD, and sex-differential risk, of families from a single, uniformly ascertained cohort. This also allows us to attempt replication of previous linkage findings.

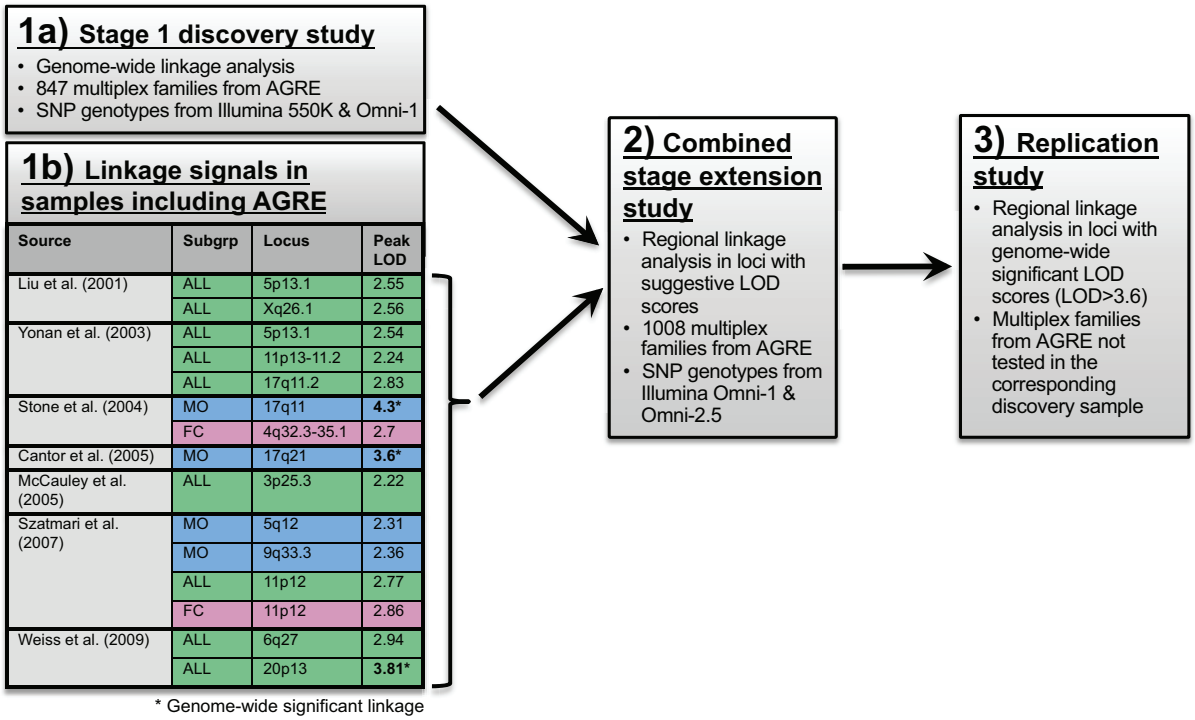


Figure 3.1: Schematic of 3-stage study design for linkage analyses
Linkage analyses were performed using a 3-stage approach. 1a) A discovery sample (Stage 1) and 1b) previous linkage results were used to identify loci of suggestive linkage (LOD≥2.2), 2) These suggestive loci were tested for linkage in an extended sample, and 3) For loci achieving genome-wide significant linkage (LOD≥3.6) in either the discovery, extended, or previously published studies, the non-overlapping subset of families was tested for independent replication of linkage.

3.3: Materials and methods

3.3.a: Subjects and genotyping

Subjects were individuals from the AGRE (Autism Genetic Resource Exchange) collection of nuclear families, including affected probands, their unaffected siblings, and their parents. Subjects provided written informed consent, or if minors, assent with agreement from their parents, to AGRE for diagnosis and blood collection. This study was approved by the Western Institutional Review Board (AGRE), the Institutional Review Board at Washington University (subject recruitment, PI: John Constantino), and by the Medical Institutional Review Board 3 at UCLA.

Individuals with study diagnoses of full autism, not quite autism, or broad spectrum disorder based on a clinician's best estimate given Autism Diagnostic Interview-Revised (ADI-R) and Autism Diagnosis Observation Schedule (ADOS) scores were considered to be affected, in accordance with the inclusive, single spectrum concept of ASD as now codified in the Diagnostic and Statistical Manual-5 (DSM-5). In the AGRE cohort, a study diagnosis of not quite autism is given to subjects who meet the age of onset criterion but score no more than one point short of the autism cutoffs in one or more symptom domains (social, communication, restricted and repetitive behavior), or subjects who meet autism cutoffs in all symptom domains, but do not meet the age of onset criterion. A study diagnosis of broad spectrum disorder is given to subjects with pervasive developmental disorders and varying levels of impairment; these subjects include individuals with conditions like PDD-NOS and Asperger's syndrome.

Probands with syndromic autism, significant dysmorphology, documented pre- or peri-natal insult, abnormal imaging or medical test, premature birth at less than 35 gestational weeks, or chromosomal abnormalities were not included; additional neuropsychiatric phenotypes in parents and unaffected siblings were not applied as inclusion or exclusion criteria. Here, chromosomal abnormalities refer to clinically relevant CNVs identified by karyotype and to recurrent CNVs such as 16p11.2 deletions and duplications independently reported to AGRE by investigators. Subjects with evidence for *de novo* missense single nucleotide variants (SNVs, N=5 cases) (Neale et al., 2012) were not excluded, due to the current uncertainty in determining the effects of missense variation on ASD risk. In instances of monozygotic multiples, only one proband was selected at random for inclusion. Subjects in the AGRE cohort include individuals of Caucasian, African-American, Asian, and Hispanic ancestry as noted by self-report and multi-dimensional scaling from genotype data; subjects were not filtered by ancestry, as the genetic

analyses used in this study (non-parametric linkage, TDT) were family-based and therefore not susceptible to the introduction of false positive results from population stratification. However, we note that including multiple ethnicities may introduce or exacerbate locus heterogeneity, which is unlikely to falsely inflate LOD scores but instead may reduce power in linkage studies (Szatmari et al., 2007).

Subjects were genotyped in two stages, using DNA purified from lymphoblastoid cell lines and obtained from the Rutgers University Cell and DNA Repository (RUCDR; Piscataway, NJ). Stage 1 consisted of individuals from 1191 AGRE families, with subjects from 941 families genotyped on the Illumina 550K genome-wide SNP array as described previously (Wang et al., 2009), and individuals from an additional 250 AGRE families typed on the Illumina Omni-1 Quad array (Illumina, San Diego, CA) in the UCLA Neuroscience Genomics Core. Stage 2 consisted of individuals from an additional 396 AGRE families, 116 of which were genotyped on the Illumina Omni-1 array, and 280 of which were typed on the Illumina Omni-2.5 array (Table 3.1). In total, the combined sample of stage 1 plus stage 2 was comprised of individuals from 1587 families, 1008 of which were multiplex and met inclusion criteria as described above.

Recorded sample identity and pedigree relationships were validated by evaluating estimations of identical by descent (IBD) allele sharing across the genome within and between families using PLINK software (Purcell et al., 2007). In cases with evidence of identity swaps, where available, genotype data from the Broad Institute (Affymetrix 5.0, [33]) and the Autism Genome Project (Illumina 1M, [37, 38]) were compared to make a final determination of identity. Subjects and SNPs with >5% missing data were excluded, and SNPs with Hardy-Weinberg equilibrium p-values <0.0000001, minor allele frequency <0.01, and >10 Mendelian errors were also excluded. Filtered data sets from the two stage 1 genotyping platforms were then

merged using PLINK to generate a stage 1 sample data set with all remaining subjects and the union of marker sets from the 550K and Omni-1 platforms for a total of 1,092,577 SNPs. To incorporate stage 2 and build the combined sample data set of all genotyped AGRE subjects, the union of filtered data from all genotyping platforms (550K, Omni-1, Omni-2.5) were merged using PLINK for a total of 1,684,432 SNPs.

3.3.b: Linkage analyses using all families

Autosomal and X chromosome markers common to all platforms and pruned to a linkage disequilibrium $r^2 \leq 0.1$ with PLINK were used for genome-wide linkage analysis (stage 1 data: 57,929 SNPs; combined stage data: 53,648 SNPs). These sets of independent SNPs were mapped to genetic positions using Rutgers Combined Linkage-Physical maps (Kong et al., 2004; Matise et al., 2007) for linkage testing. Non-parametric, multipoint linkage was performed genome-wide on all stage 1 multiplex families meeting inclusion criteria (N=847 families) using Merlin (autosomes) and Minx (X chromosome, executable option within Merlin) (Abecasis, Cherny, Cookson, & Cardon, 2002) which applies the Kong and Cox linear model (Kong & Cox, 1997) to test for small increases in allele sharing across a large sample of families.

Linked regions identified in the stage 1 sample or by previous reports (discovery samples) from non-parametric linkage analyses of AGRE samples using the "broad" affection status criteria were then identified for further evaluation by defining the 2-LOD (logarithm of odds) interval surrounding suggestive (LOD>2.2) and significant (LOD>3.6) signals. Since the precise physical position of linkage peaks may vary between studies using marker sets of differing density and information content (see Cantor et al., 2005 for an example), this 2-LOD interval is intended to be inclusive and to encompass any underlying variability in the location of

previous signals. Where there were discrepancies between the genetic maps from the original studies and our data, the span of the interval was anchored on the current physical position (hg19) of the peak marker. Linkage signals from the study by Szatmari and colleagues (2007) were reported as $Z_{|r}$ scores, which we converted to LOD scores by: $LOD = Z_{|r}^2/(2*\ln(10))$ (A. Kong & Cox, 1997). Regions with suggestive linkage evidence from stage 1 or previous reports were tested for linkage in the combined sample. Regions at which the discovery sample showed genome-wide significant linkage (LOD>3.6) were then evaluated for replication ($p<0.01$) within a 2-LOD interval from the peak LOD by testing only those independent families not previously evaluated in the corresponding discovery analysis.

We applied this 1) discovery, 2) extension, and 3) replication approach according to guidelines for linkage analyses of complex traits outlined by Lander and Kruglyak (Lander & Kruglyak, 1995). As preliminary linkage analyses for complex traits have yielded suggestive signals, extending sample sizes by adding pedigrees may allow these loci to reach genome-wide significance. Once a genome-wide significant signal has been identified, it may then be tested for independent replication in a new sample. The platform and stage at which a family was genotyped determined whether they were included in the discovery (stage 1) or the extension (combined stage) stages (Figure 3.1).

3.3.c: Sex-stratified linkage

To identify sex-differential ASD risk loci, we stratified the genotyped, multiplex families into two groups based on the sex of their affected children: male-only (MO, no affected daughters), and female-containing (FC, at least one affected daughter) (Stone et al., 2004); the number of affected-female-only multiplex families enrolled in AGRE is currently too low to

analyze these families as their own subgroup. In keeping with earlier sex-stratified linkage analyses of AGRE families (Cantor et al., 2005; Stone et al., 2004; Szatmari et al., 2007) and to maintain workable sample sizes, all families with two or more affected members were assigned by this simple criterion to either the MO or FC subset, irrespective of the total affected family members or the presence of unaffected siblings of the opposite sex. While it can be argued that MO families with three or more affected brothers or with at least one unaffected sister are more likely to carry truly male-specific risk variants (not penetrant in females) than MO families with fewer affected brothers or no unaffected sisters, the number of available AGRE families fitting these criteria is quite low. Specifically, 52 MO families include three or more affected brothers, and 150 MO families have recorded unaffected sisters. To avoid further restricting the sample size of our subsets, we define the MO and FC subsets simply by the sex of the two or more affected siblings, under the assumption that the MO and FC subsets are enriched for male-specific and female-affecting risk variants, respectively, as compared with the full, non-stratified sample. By this definition, the stage 1 sample consisted of 487 MO (61%) and 314 FC (39%) families, and the combined sample consisted of 602 MO (60%) and 406 FC (40%) families (Table 3.1).

Table 3.1: Genotyped families and cases from AGRE

	Stage	Platform		Total	Multiplex	MO	FC
Combined sample	Stage 1	Illumina 550K	Families	941	704	431	273
			Cases (females)	1729 (368)	1481 (319)	886 (0)	595 (319)
		Illumina Omni-1	Families	250+29*	128	77	51
			Cases (females)	381 (78)	232 (52)	134 (0)	98 (52)
	Stage 2	Illumina Omni-1	Families	116+11*	27	9	18
			Cases (females)	141 (39)	51 (20)	19 (0)	32 (20)
		Illumina Omni-2.5	Families	280+3*	151	89	62
			Cases (females)	478 (102)	347 (82)	188 (0)	159 (82)

*Additional members from families partially genotyped at earlier stage

Linkage analyses were performed separately in the two subgroups in three steps as described above for the non-stratified sample: 1) genome-wide non-parametric linkage analysis of MO and FC subgroups from the stage 1 sample, 2) regional linkage analysis within suggestive (LOD>2.2) MO- or FC-specific linkage peaks from discovery studies (Cantor et al., 2005; Stone et al., 2004; Szatmari et al., 2007) using the corresponding MO or FC subgroup from the combined sample, 3) regional linkage analysis within peaks of genome-wide significant LOD (>3.6) in the independent portion of the corresponding MO or FC subgroup from the combined sample who were not previously tested in the discovery study.

Additionally, to assess the statistical significance of sex-differential linkage signals, a randomization test of 10,000 subsets of 487 families (matching stage 1 MO family N) and 314 (stage 1 FC family N) were analyzed for linkage across chromosomes 1, 4, 6, and 8, where subgroup-specific suggestive linkage peaks (LOD ≥ 2.2) were observed. For the combined sample, random subsets of 602 (MO family N) were analyzed across chromosome 1, the only chromosome on which a sex-differential signal surpassed LOD 2.2 in the combined sample. On a marker-by-marker basis, the LOD score from each of the 10,000 random trials was compared to the results from the corresponding original, subset-specific scan, and the fraction of random trials for which $LOD_{random} > LOD_{original}$ was taken as the empirical p-value for stratification, reflecting the frequency with which the observed LOD magnitude would occur under the null hypothesis of no linkage.

3.3.d: Imputation

To further improve genotype coverage within linked regions for fine-scale association testing, imputation was performed separately by data set, as defined by genotyping platform and

data collection stage, using IMPUTE2 (Howie, Donnelly, & Marchini, 2009) and a cosmopolitan reference panel from the 1000 Genomes Project (Abecasis et al., 2010). Imputed SNPs from each data set were then merged using GTOOL (Marchini & Howie, 2010), and SNPTEST (Marchini & Howie, 2010) was used to generate summary statistics for the merged set. Data were filtered to SNPs with an IMPUTE2 quality score ≥0.5, missing data in ≤5% of subjects, minor allele frequency ≥1%, and Hardy Weinberg p≥0.0000001. The final data set included 5,814,564 autosomal SNPs.

3.3.e: Linkage-directed association testing

Imputed SNPs within the 2-LOD intervals surrounding linkage peaks exceeding the suggestive threshold (LOD>2.2) in either the stage 1 or combined sample were tested for association with ASD affection status in the family group corresponding to the linkage peak (ALL, MO, or FC) using a transmission disequilibrium test (TDT) with adaptive gene-dropping permutations (PLINK). Gene dropping assignments are applied consistently across siblings to control for linkage, therefore appropriately treating trios from multiplex families as non-independent (Purcell et al., 2007).

In regions of MO or FC linkage signal, extended families without multiple affected siblings, but instead with multiple affected cousins, were additionally included in the association analysis (total MO N=606, including 4 additional extended families; total FC N=407, including 1 additional extended family). In regions of linkage signal from all families (ALL), an additional 5 extended families and 508 families with only one genotyped, affected individual were included in the association analysis (total ALL N=1521 families). All additional families met inclusion criteria as described previously. Association p-values were adjusted for multiple testing

according to the number of independent SNPs within each region, as defined by a pairwise linkage disequilibrium $r^2<0.3$ (PLINK).

3.4: Results

By testing the largest available, uniformly ascertained ASD family sample (n=1008 multiplex families) for linkage, we aimed to identify novel ASD risk loci not identified by the previous, smaller analyses of AGRE samples, and/or to confirm the loci reported by these analyses. As genotype data were collected in two stages, we use the stage 1 sample to identify linked and sex-differential risk loci genome-wide, and the combined sample (union of stages 1 and 2) to evaluate identified loci from stage 1 and earlier reports (discovery studies) for confirmation of linkage in the largest available AGRE genotype data set.

3.4.a: Linkage in all families

Non-parametric, genome-wide linkage analysis for ASD affection status in all stage 1 multiplex families (n=847) identified four genomic regions with a peak LOD score >2.2 (Figure 3.2, Table 3.2), the threshold for suggestive linkage for a complex trait when allele sharing is tested in sibling pairs (Lander & Kruglyak, 1995). We observed the highest LOD score at chromosome 6q27, with $LOD_{ALL.St1}$=3.22 at rs4708676 (190.611 cM). A 2-LOD interval from this peak SNP spans 18.2 Mb, 31.6 cM, and 100 RefSeq genes, and the peak SNP is 75 kb upstream from gene *FRMD1* [GenBank:NM_024919] (0.4-LOD drop from peak), a gene with a role in immune function and significantly associated with IL-2 secretion (Kennedy et al., 2012). All other linkage regions yielded peak LOD >2.2 and are located at chromosomes 4q13.1 ($LOD_{ALL.St1}$=2.3), 8p21.2 ($LOD_{ALL.St1}$=2.55), and 8q13.2 ($LOD_{ALL.St1}$=2.5).

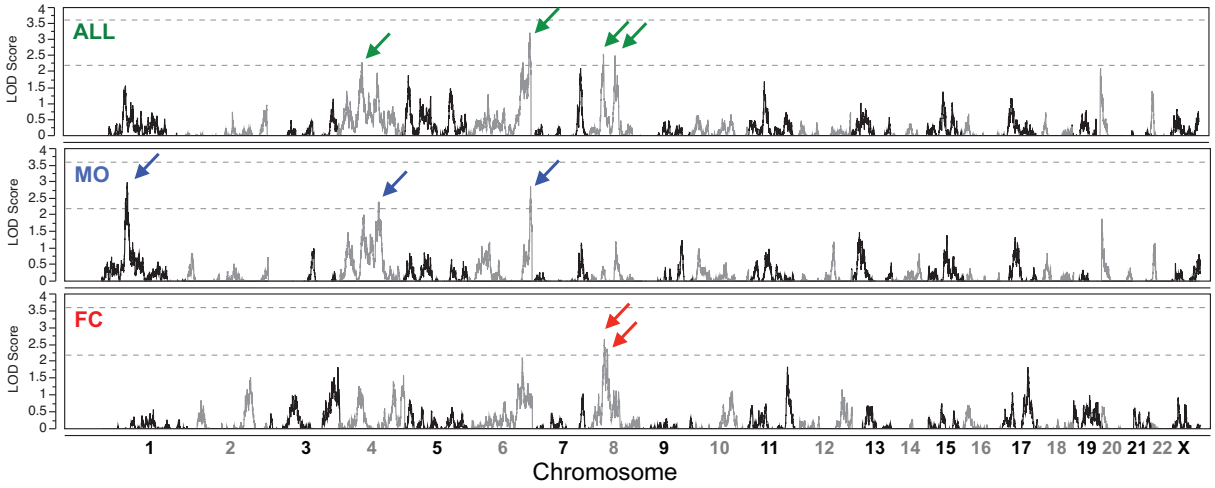


Figure 3.2: Genome-wide linkage, stage 1 sample
LOD scores from genome-wide, non-parametric, multipoint linkage analysis for ASD affection status in the stage 1 sample are plotted for autosomes (Merlin) and chromosome X (Minx). Top panel: all multiplex families; middle: male-only families, "MO"; bottom: female-containing families, "FC". Dashed lines mark LOD thresholds 2.2 for suggestive, and 3.6 for significant, linkage [39]. Arrows note signals with LOD≥2.2.

To determine if the signals observed in the stage 1 sample and in previous AGRE studies could be improved or replicated in a larger sample size, we next carried out an extension study by testing these regions for linkage in the combined sample of 1008 multiplex families (union of stage 1 and stage 2 samples). Non-parametric linkage analysis of the combined sample within the regions of interest confirmed three loci above the suggestive threshold of LOD>2.2: 6q27 ($LOD_{ALL.Com}$=2.50), 8q13.2 ($LOD_{ALL.Com}$=2.82), and 20p13 ($LOD_{ALL.Com}$=3.02; Figure 3.3). Of these three loci, only the signal at 8q13.2 increased in the combined sample. Combined sample linkage on other chromosomes of interest is shown in Figure 3.4.

The strongest signal that we observed from the combined sample, and the only signal that surpassed a LOD score of 3.0, is at 20p13. This locus had been identified previously in study that combined 753 AGRE families with other families from the NIMH repository (Weiss et al., 2009). Since 665 of the families in that analysis overlapped with the current sample, we formally tested for replication of this locus by analyzing only those families unique to our study (n=343

Table 3.2: Summary of significant and suggestive linkage peaks from AGRE samples

Discovery sample					Combined sample (Stage 1 + Stage 2)			
Source	**Number of families**	**Group**	**Locus***	**Peak LOD**	**No. of families**	**Peak LOD**	**Peak p-value**	**Peak SNP**
Liu et al. (2001)	110	ALL	5p13.1	2.55	1008	0.99	0.0163	rs6884342
	110	ALL	Xq26.1	2.56	1008	1.12	0.0116	rs12557711
Yonan et al. (2003)	345	ALL	5p13.1	2.54	1008	0.99	0.0163	rs6884342
	345	ALL	11p13-11.2	2.24	1008	0.92	0.0199	rs2984699
	345	ALL	17q11.2	2.83	1008	0.53	0.0592	rs1382779
Stone et al. (2004)	148	MO	17q11	**4.3**	602	0.42	0.0834	rs4795708
	109	FC	4q32.3-35.1	2.7	406	1.10	0.0121	rs1717072
Cantor et al. (2005)	196	MO	17q21	**3.6**	602	0.47	0.0706	rs1877032
McCauley et al. (2005)	158 (85 AGRE)	ALL	3p25.3	2.22	1008	0.04	0.3344	rs1400207
Szatmari et al. (2007)	741 (211 AGRE)	MO	5q12	2.31	602	1.42	5.34e-3	rs706725
	741 (211 AGRE)	MO	9q33.3	2.36	602	1.47	4.64e-3	rs204169
	1181 (387 AGRE)	ALL	11p12	2.77	1008	0.92	0.0199	rs2984699
	440 (176 AGRE)	FC	11p12	2.86	406	0.38	0.0925	rs404977
Weiss et al. (2009)	904 (753 AGRE)	ALL	6q27	2.94	1008	2.50†	3.43e-4	rs6931082
	904 (753 AGRE)	ALL	20p13	**3.81**	1008	3.02†	9.55e-5	rs6139007
Current study Stage 1 sample	487	MO	1p31.3	2.98	602	2.55†	3.05e-4	rs7521242
	847	ALL	4q13.1	2.30	1008	2.14	8.37e-4	rs1483288
	487	MO	4q26	2.41	602	1.28	7.60e-3	rs2196712
	847	ALL	6q27	3.22	1008	2.50†	3.43e-4	rs6931082
	487	MO	6q27	2.86	602	2.07	1.02e-3	rs960145
	314	FC	8p21.2	2.67	406	1.42	5.25e-3	rs7001120
	847	ALL	8p21.2	2.55	1008	2.18	7.60e-4	rs13257637
	314	FC	8p12	2.37	406	1.34	6.51e-3	rs2976525
	847	ALL	8q13.2	2.50	1008	2.82†	1.58e-4	rs4738003

Bold type indicates LOD scores passing genome-wide significance thresholds of LOD≥3.6 for discovery studies, †LOD scores passing suggestive linkage thresholds of LOD≥2.2 in the combined sample extension study.

multiplex families). We identified a linkage peak within the 20p13 region of interest with a p-value of 0.0076 at rs214828, thus meeting the significance threshold for independent replication of linkage for a complex trait ($p<0.01$) (Lander & Kruglyak, 1995). An additional five nearby

SNPs also had p-values of less than 0.01 in this independent family set. From the peak $LOD_{ALL.Com}$ of 3.02 at SNP rs6139007 (1.613 cM) in the combined sample, a 2-LOD drop support interval at this locus spans 4.8 Mb, 14.2 cM, and 87 RefSeq genes (Figure 3.3). The combined sample peak SNP is located just 500 bp upstream from the transcription start site of *TRIB3* [GenBank:NM_021158] (0.1-LOD drop from peak), which encodes a regulator of *AKT1* [GenBank:NG_012188] and is expressed mainly in pancreas, bone marrow, and leukocytes (Du, Herzig, Kulkarni, & Montminy, 2003; Kiss-Toth et al., 2004). The independent sample peak SNP, rs214828 at 8.403 cM, is intronic to gene *TGM3* [GenBank:NM_003245] (0-LOD drop from peak), which encodes a calcium-dependent peptide cross-linking enzyme (Kim et al., 1994).

While signals at 6q27 and 8q13.2 did not achieve genome-wide significance in either stage of analysis, the consistency of the signal at these loci across the discovery and combined samples suggests these regions may harbor ASD risk variants; each encompasses promising candidate genes, including *SULF1* [GenBank:NM_001128206] (0-LOD drop from peak at 8q13.2) located directly under the 8q13.2 linkage peak and whose protein product interacts with growth factors and cytokines in cell signaling (Dai et al., 2005), and *PARK2* [GenBank:NG_008289] and *RPS6KA2* [GenBank:NM_021135] (1.7- and 0.6-LOD drop, respectively, from peak at 6q27), which are both located within rare CNVs identified in ASD cases (Glessner et al., 2009; Marshall et al., 2008; Scheuerle & Wilson, 2011).

3.4.b: Sex-stratified linkage

To identify sex-differential ASD risk loci, we analyzed the multiplex families in two separate groups according to the sex of the affected children in each family: male-only (MO) and

female-containing (FC) (see Methods) (Stone et al., 2004). This sex stratification approach has been applied only twice in earlier analyses of exclusively AGRE families, in subgroups only one third as large as our stage 1 MO and FC subgroups (Cantor et al., 2005; Stone et al., 2004). These earlier analyses identified and replicated a genome-wide significant signal at 17q11-q21 in the MO subgroup, which has not been subsequently replicated in larger studies (Szatmari et al., 2007). Using our larger sample, we aimed to identify additional sex-differential risk loci from both the MO and FC subgroups.

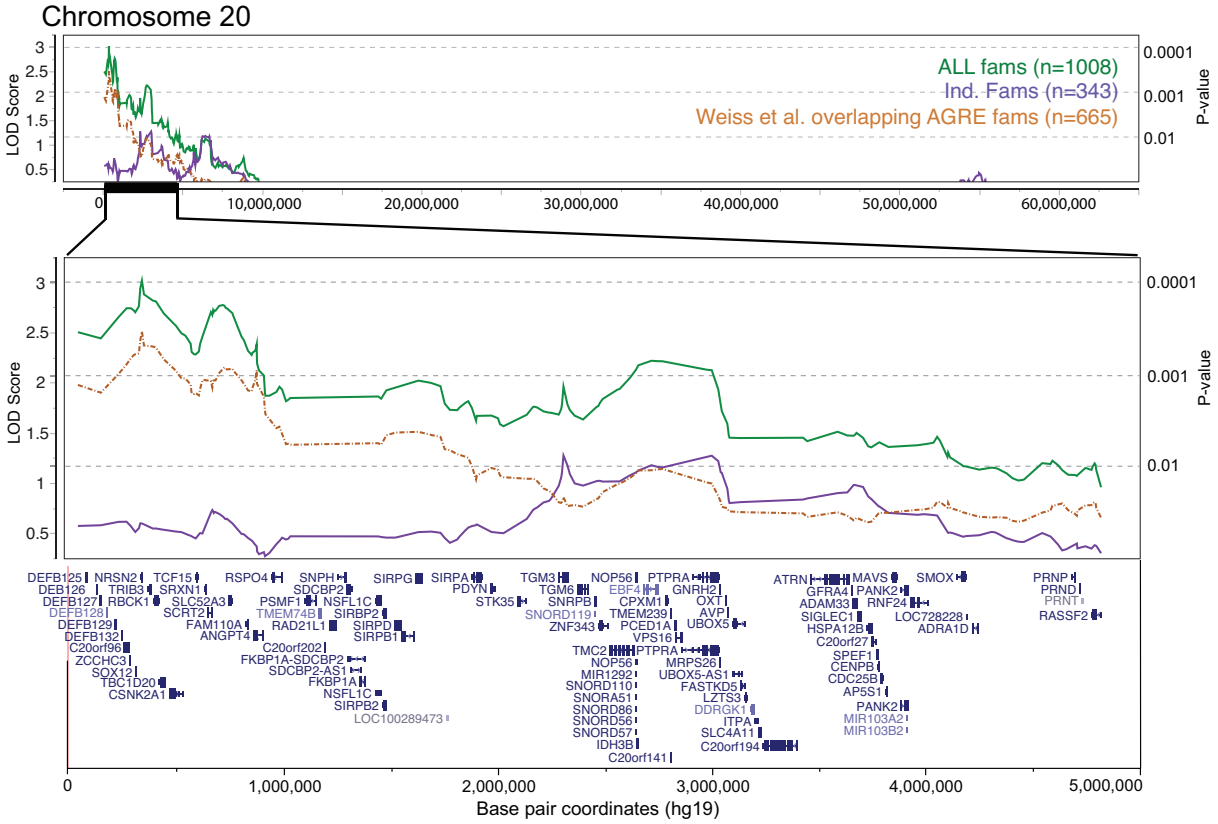


Figure 3.3: Independent replication of genome-wide significant linkage at 20p13
Genome-wide significant linkage signal at 20p13 from the combined sample (green, solid line), AGRE families analyzed by Weiss et al. (2009) (orange, dashed line), and all AGRE families *not* previously analyzed by Weiss et al. (purple, solid line). Top: linkage across the full chromosome 20; middle: linkage across a 2-LOD interval from the peak LOD; bottom: RefSeq gene alignment in the 2-LOD interval. Dashed lines mark LOD thresholds corresponding to linkage p-values of 0.01, 0.001, and 0.0001.

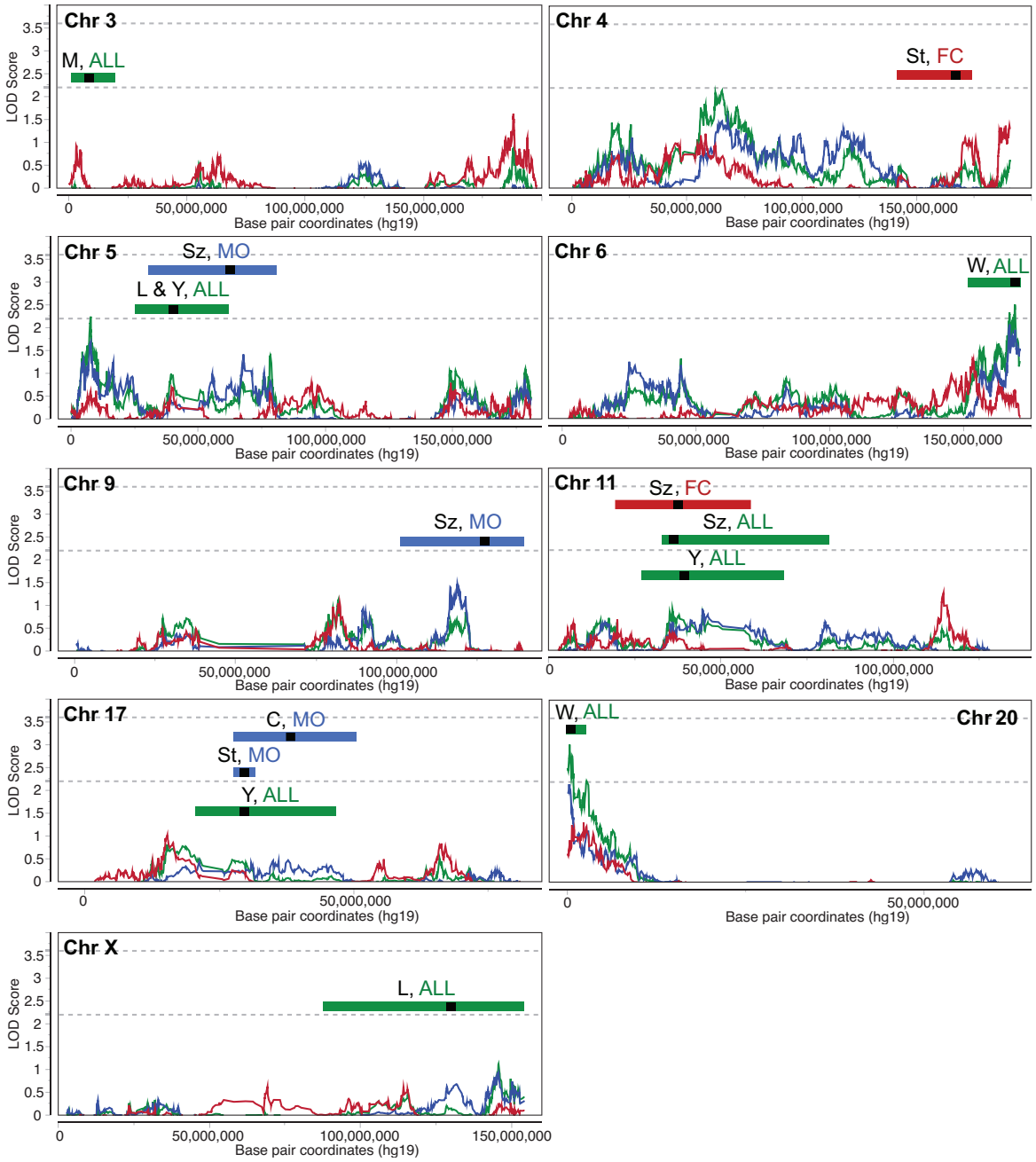


Figure 3.4: Linkage in regions of interest from previous studies, combined sample
LOD scores from non-parametric, multipoint linkage analysis for ASD affection status in the combined sample (Merlin) are plotted for chromosomes with suggestive linkage peaks (LOD>2.2) from previous studies. Colored bars and black marks indicate the spans and peaks of linkage regions of interest, respectively; text indicates the linkage region source (L= Liu et al., 2001; Y=Yonan et al., 2003; St=Stone et al., 2004; C=Cantor et al., 2005; M=McCauley et al., 2005; Sz=Szatmari et al., 2007; W=Weiss et al., 2009; green and ALL=all multiplex families; blue and MO=male-only families; red and FC=female-containing families). Dashed lines mark LOD thresholds of 2.2 for suggestive, and 3.6 for significant, linkage

Table 3.3: Sex subset-specific linkage signals

Locus	Subgroup	Stage 1 sample peak LOD	Combined sample peak LOD	Gene at suggestive linkage peak	Sex-differential linkage empirical p-value
1p31.3	MO	2.98	-----	NFIA	**p<0.01**
	MO	-----	2.55	NFIA	**p<0.0005**
4q26	MO	2.41	1.28	SYNPO2	NS
6q27	MO	2.86	2.07	MLLT4	NS
8p21.2	FC	2.67	1.42	EBF2	**p<0.05**
8p12	FC	2.37	1.34	NRG1	**p<0.005**

Bold type indicates significant sex-differential linkage signals with empirical p≤0.05.

Separate non-parametric, genome-wide linkage analyses for ASD affection status in the stage 1 MO (n=487 families) and FC (n=314 families) subgroups identified five loci with LOD scores >2.2 (Table 3.3, Figure 3.2), two of which overlap peaks from all families (6q27, 8p21.2) and three that are suggestive only in either the MO or FC subgroup (1p31.3, 4q26, 8p12).

We observed the highest LOD score for the MO subset at chromosome 1p31.3, with $LOD_{MO.St1}$=2.98 at rs7521242 (92.905 cM); a 2-LOD interval from this peak SNP spans 10.0 Mb, 13.7 cM, and 48 RefSeq genes. Analysis of this locus in the MO subgroup from the combined sample identified a smaller, but still suggestive, peak $LOD_{MO.Com}$=2.55 also at rs7521242 (Figure 3.5A). In both samples, this MO-specific peak is centered on gene *NFIA* [GenBank:NG_011787] (0-LOD drop from peak), which is expressed in the central nervous system and plays a significant role in glial cell fate determination and in normal development of the corpus callosum (das Neves et al., 1999; Deneen et al., 2006). Exome sequencing has also identified a *de novo,* non-synonymous, loss of function SNV in an autistic subject in this gene, although the SNV carrier is female (Iossifov et al., 2012).

From the FC stage 1 subset, we observed the highest LOD score at chromosome 8p21.2, with $LOD_{FC.St1}$=2.67 at rs10111167 (46.967 cM); a 2-LOD interval from this peak SNP spans 6.4Mb, 9.9 cM, and 73 RefSeq genes (Figure 3.5C). This FC-specific peak is centered on gene

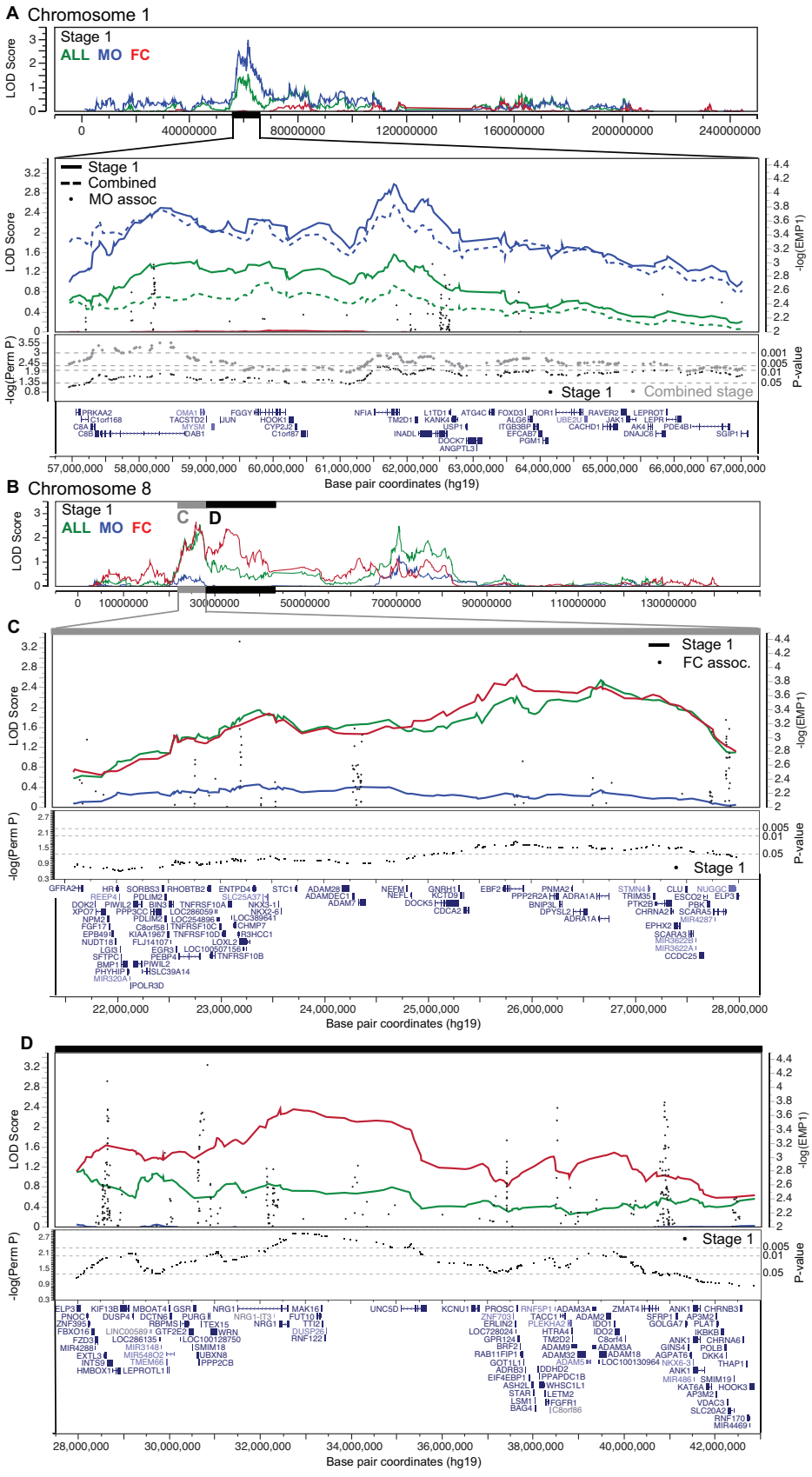


Figure 3.5: Significant sex-differential linkage peaks
Regions of sex subgroup-specific suggestive linkage (LOD>2.2) with empirically significant signal enrichment from randomization testing **A)** 1p31.3, MO; top: linkage across full chr. 1 from all stage 1 family groups; upper-middle: linkage from all stage 1 (solid lines) and combined sample (dashed lines) family groups and association signal from TDT (black points, EMP1 is empirical p-value from TDT) in MO families across 2-LOD interval from peak LOD; lower-middle: p-values from test of linkage subgroup specificity (black = stage 1, gray = combined), dotted lines indicate p-value thresholds of 0.05, 0.01, and 0.005; bottom: RefSeq gene alignment in interval. **B)** Linkage across full chromosome 8 from all stage 1 family groups. **C)** 8p21.2, FC, **D)** 8p12, FC; for each, top: linkage from all stage 1 family groups and TDT association in FC families across 2-LOD interval from peak LOD; middle: empirical p-values from stratification permutation tests; bottom, RefSeq gene alignment in 2-LOD interval.

EBF2 [GenBank:NG_030344] (0-LOD drop from peak), which encodes a transcription factor that may act alongside WNT1 [GenBank:NG_033141] to regulate cellular differentiation during development (Kieslinger et al., 2005). However, LOD scores in this region from the combined sample FC subgroup did not reach the suggestive LOD threshold of 2.2. Similarly, the other sex subgroup-specific peaks at chromosomes 6q27 ($LOD_{MO.St1}$=2.86), 4q26 ($LOD_{MO.St1}$=2.41), and 8p12 ($LOD_{FC.St1}$=2.37), only exceeded LOD of 2.2 in the stage 1 sample (Table 3.2, Table 3.3).

We next sought to calculate the likelihood of observing the MO and FC signals that surpassed (1p31.3, 4q26, 8p21.2, 8p12), or approached (6q27) the magnitude of the LOD signal from all families by chance. To assess the significance of this signal enrichment in the sex-defined subsets, we performed a randomization test for linkage in 10,000 N-matched subsets of families from the full multiplex family cohort. The empirical p-value indicates subset-specific linkage enrichment (empirical $p<0.05$ at peak SNP) at chromosomes 1p31.3 (MO) in both the stage 1 and the combined samples, and at 8p21.2 (FC), and 8p12 (FC) in the stage 1 sample (Table 3.3, Figure 3.5). These results suggest that the stratification of cohorts by proband sex can reveal sex-differential ASD risk loci. However, only the MO locus at 1p31.3 was supported in the combined sample.

Aside from the signal identified by Weiss and colleagues (2009) at 20p13 in a non-sex-stratified sample, the only other locus at which genome-wide significant linkage was previously observed was in a MO subset at 17q11 (Stone et al., 2004). This signal was subsequently replicated in an additional MO subset of AGRE families (Cantor et al., 2005). In an attempt to further replicate the signal at this locus, we tested the 407 MO families from the combined stage that were not included in either previous analysis. The most significant linkage p-value that we observe within the 2-LOD drop interval from the peak signal observed by Cantor and colleagues

(2005) is 0.1003 at rs2014209; thus, we do not independently replicate the MO signal at 17q in this sample.

3.4.c: Linkage-directed association

We next tested for association with common genotyped and imputed SNPs within the 2-LOD intervals around each linkage peak with LOD>2.2 in the stage 1 or the combined sample. For regions with suggestive signals from the MO or FC subsets, association testing was run on the corresponding subset from the combined stage, with the addition of families with multiple affected cousins instead of multiple affected siblings. For regions with suggestive signals from all multiplex families (ALL), association testing was run on all multiplex families from the combined stage with the addition of 508 families with only a single genotyped, affected member. While we observe some clusters of SNPs that approach significance, we did not identify any SNP that survived multiple testing correction for the number of independent SNPs within each region, defined by pairwise LD $r^2<0.3$. We highlight the top association signals here for interest's sake, although we emphasize that none pass our correction for multiple comparisons.

The strongest unadjusted association signal within any linkage region occurred in the ALL peak on chromosome 4q13.1 at rs115667468 with a p-value of 2.997×10^{-5} (after regional correction, p=0.154). This associated SNP is located 7.8 Mb from the SNP at the linkage peak and is intronic to gene *NPFFR2* [GenBank:NM_004885], and the minor allele T was found to be under-transmitted to affected offspring from ALL families (odds ratio = 0.4143). In contrast, the strongest corrected association signal occurred in the FC peak on chromosome 8p21.2 at rs78485638 with a corrected p-value of 0.052 (unadjusted $p=4.206 \times 10^{-5}$; Figure 3.5D). This SNP is located 2.7 Mb from the SNP at the linkage peak, is intronic to gene *LOXL2*

[GenBank:NG_002318], and the minor allele T was found to be under-transmitted to affected offspring from FC families (odds ratio = 0.4512).

3.5: Discussion

To identify and support genomic loci likely to contain variants contributing to ASD risk in multiplex families from the AGRE collection, we performed linkage analyses in all and sex-stratified subsets of multiplex families followed by targeted association testing in 1521 families (1008 multiplex) from the AGRE cohort. The strongest linkage signal that we identified was on chromosome 20p13, which exceeded a LOD score of 3.0 in our combined sample of 1008 multiplex families. At 20p13, we also replicated in an independent sample from the same AGRE cohort a previous report of significant linkage at this locus (Weiss et al., 2009). Analyses of sex-defined family subgroups and randomization testing for signals from these subgroups identified a locus at chromosome 1p31.3 that showed significant linkage in the MO subgroup in both the stage 1 and combined samples, and loci at chromosomes 8p21.2 and 8p12 that showed significant linkage in the FC subgroup only in the stage 1 sample. No genotyped or imputed common SNPs within any linked region proved to be significantly associated with ASD, an observation consistent with a model where the influence of multiple loci of very small effect size, or of rare variants, contributes to ASD.

The linkage signal at 20p13 is especially noteworthy as it was the most significant signal that we observed in our combined sample with a $LOD_{All.Com}$ of 3.02, and it also meets the criterion for an independent replication of genome-wide significant linkage (Lander & Kruglyak, 1995). In addition, as both this study and the study by Weiss and colleagues (2009) were conducted using the AGRE collection, a sample that has been uniformly ascertained and

evaluated over time, this is a clear replication. A 2-LOD interval from the combined sample peak spans such potential risk candidate genes as *NRSN2* [GenBank:NM_024958] (0.3-LOD drop from peak), a gene expressed throughout the cerebral cortex, thalamus, hypothalamus, and in Purkinje cells (Nakanishi et al., 2006), and *CSNK2A1* [GenBank:NG_011970] (0.5-LOD drop from peak), which encodes a protein involved in the regulation of circadian rhythms (Lin et al., 2002).

For nearly all linked regions identified in the discovery samples, including 20p13, LOD scores decreased when analyzed in the combined sample despite the increased sample size, consistent with previous observations of genetic heterogeneity (Huang & Vieland, 2001; Vieland, Wang, & Huang, 2001; Yonan, Palmer, et al., 2003). However, it is interesting to note that those signals that confirmed or replicated in the combined or independent sample were initially observed by tests of relatively large discovery samples. This is because, like association testing, a very large number of small pedigrees are critical to identify robust linkage signals (Sham, Lin, Zhao, & Curtis, 2000). In smaller samples, signal fluctuations between analyses are caused by extreme heterogeneity of genetic risk loci for ASD (Bartlett, Goedken, & Vieland, 2005; Huang & Vieland, 2001; Vieland et al., 2001), such that each analysis identifies different, but potentially true, risk loci. Since linkage peaks represent deviations from expected proportions of affected family members sharing two, one, or zero alleles IBD at a particular locus, linkage analyses for complex traits are particularly sensitive to the composition of families included in any one test. If, as projected, there are close to 1000 risk genes for ASD (Sanders et al., 2012), then the chances that a sufficient proportion of families in current sample sizes share the same risk locus are low.

We recognize that the power we have to detect linkage in our analyses is for complete linkage with loci of modest genetic effects on ASD risk (Sham et al., 2000), and so it is not surprising that we do not observe genome-wide significant linkage in either the stage 1 or extended, combined stage samples, and that most suggestive signals fail to increase in the extended sample. However, we emphasize that only now are we approaching the sample sizes necessary to detect significant linkage in the face of locus heterogeneity in cohorts comprised of small pedigrees, and that any possible reduction of locus heterogeneity, e.g. by testing families ascertained by the same cohort as is done here, will be key to identify replicable linkage signals.

Rational stratification of cohorts into subgroups based on a shared trait may also facilitate the discovery of risk loci by increasing the relative homogeneity of specific genetic risk factor(s) in the subgroup. Here, we stratified our sample by the sex of the probands within each family to identify loci with sex-differential relationships with ASD risk, and found significant risk loci at chromosome 1p31.3 (MO), 8p21.2 (FC), and 8p12 (FC). At 1p31.3 and 8p12, linkage signals from the MO and FC subsets, respectively, were significantly stronger than the signal from the full, non-stratified cohort, suggesting that this sex-based stratification approach can indeed reduce the genetic heterogeneity within each subgroup that would otherwise obscure signals at these loci.

We note that these loci should not be interpreted as simply sex-specific. Namely, although we cannot say that the MO group is perfectly restricted to families who carry solely male-specific variants and therefore only have, and would ever have, affected male children, we do assume that the MO subgroup is substantially enriched for families who carry risk variants that are more penetrant in males. Thus, signals specific to the MO subset are more likely to be male-specific. In contrast, since the FC group includes affected brothers of autistic girls, FC

signals are best interpreted not as female-specific, but as female-affecting risk loci, in accord with a hypothesis that only a subset of ASD risk loci are penetrant in females (Neale et al., 2012; O'Roak et al., 2012; Robinson et al., 2013; Sanders et al., 2011; Sanders et al., 2012; Sebat et al., 2007).

The significant MO and FC signals identified here implicate regions containing promising candidate genes that warrant further exploration by targeted re-sequencing (Brkanac et al., 2009; Ng, Nickerson, Bamshad, & Shendure, 2010; Nikopoulos et al., 2010; Rehman et al., 2010). The MO peak at 1p31.3 is located directly over *NFIA* [GenBank:NG_011787], whose gene product has transcription factor activity and has been implicated in central nervous system development (das Neves et al., 1999; Deneen et al., 2006). Rare deletions encompassing this gene have been identified in subjects with ASD (Mikhail et al., 2011), as well as *de novo* mutations (Iossifov et al., 2012). The FC peak at 8p21.2 spans several candidates, including *STC1* [GenBank:NG_029711] (1.1-LOD drop from peak), which encodes a glycoprotein regulated by calcium that may act to protect neurons from ischemia and hypoxia (Zhang et al., 2000), and neurofilament genes *NEFM* [GenBank:NG_008388] and *NEFL* [GenBank:NG_008492] (both 1.1-LOD drop from peak) whose products likely function in transport to neuronal projections (Brownlees et al., 2002). Potentially relevant to sex-differential risk, *GNRH1* [GenBank:NG_016457] (0.7-LOD drop from peak) is also located within this linkage region, and mutations in this gene are likely to affect gonadal function (Bouligand et al., 2009), perhaps differentially modulating downstream manifestation of ASD risk factors in males and females. The neighboring FC peak at chromosome 8p12 is located directly over *NRG1* [GenBank:NG_012005] (0-LOD drop from peak), a known schizophrenia risk gene (Stefansson et al., 2003; Stefansson et al., 2002).

A similar stratification approach for the identification of male-specific and female-affecting ASD risk loci has been successfully applied previously by Stone and colleagues (2004) to an early iteration of the AGRE cohort, as well as by Szatmari et al. (2007) to a sample that included a subset of AGRE families, and Lamb et al. (2005) and Schellenberg et al. (2006) to other ASD family samples. No convergence in linkage signals was observed across these studies. We predicted that our ability to identify novel, sex-differential ASD risk loci in the present study would be aided by a greater than three-fold increase in subjects exclusively from AGRE as compared to those used by Stone and colleagues (2004) and by the increase in coverage afforded by dense SNP data in lieu of several hundred microsatellite markers. However, the results from our sex-stratified analyses do not reach genome-wide significance and also do not align with findings from earlier sex stratification analyses, including Stone et al. (2004) who reported a linkage peak at 17q11-q21 in MO families that was subsequently replicated (Cantor et al., 2005). This variability between studies is again consistent with extreme risk locus heterogeneity in ASD (Vieland, 2006, 2011), with each analysis of a different combination of ASD families identifying different linked regions.

Under the assumption of genetic heterogeneity, for complex traits such as ASD, it is possible that the various loci identified by analyses of different family sets flag true sites for ASD risk in a proportion of the families tested. To pursue this, linked loci will need to be investigated more closely to identify the precise variants that effectively increase ASD risk. This has so far proven challenging, as association testing of densely mapped common SNPs within linkage peaks has failed to definitively identify risk variants, both in the present study and in previous work (Weiss et al., 2009).

Although our analyses are likely underpowered to identify regionally significant associations with common variants of small effect size, it is alternatively possible that rare variants, not explicitly tested here, contribute to the heritable component of ASD risk. For example, rare variants, shared between siblings but private to each nuclear family, may cluster in the same gene or set of genes. Since rare variants are less likely to be tagged by common SNPs, they should be more readily localized by allele-agnostic linkage analyses than association testing. As in gene discovery studies of sporadic ASD cases, sequencing of functional genomic features in linked regions will be necessary to identify rare variants and evaluate their role in familial ASD risk. In either case, larger family-based cohorts will be needed to improve power.

3.6: Conclusions

We conclude that the use of linkage analyses in multiplex family cohorts has complementary utility to genome-wide association studies for the investigation of the familial, inherited contribution to ASD risk. This is especially the case in the context of rare variants in human disease (Barzilai et al., 2003; Yokoyama et al., 2011). Additionally, the use of a sex stratification approach facilitates the identification of risk loci that are differentially associated with ASD in families with autistic sons versus daughters. However, further work is needed to determine which gene(s) or genetic features within linked regions, especially at chromosome 20p13, replicated here, harbor the variants responsible for increasing familial and sex-differential genetic risk for ASD. Exploring this in detail via targeted sequencing in large cohorts will be necessary to elucidate the common versus rare genetic contributions to ASD.

CHAPTER 4:

Strategies for the identification of functional autism risk variants in linkage regions

4.1: Linkage signals in autism spectrum disorders

The earliest evidence for a genetic contribution to the cause of autism spectrum disorders (ASDs) came from twin (Bailey et al., 1995; Folstein & Rutter, 1977), and subsequently family studies (Bolton et al., 1994; Rutter, 1968), which showed ASD to be a highly heritable condition. These findings suggested that studies of families with multiple autistic members would be an informative approach for identifying the gene responsible for ASD. Linkage analyses, for identifying genomic regions shared by affected individuals within and between families, had proven successful for identifying genes involved in other disorders with single-locus, "Mendelian," inheritance patterns, but results from ASD families have, with a few exceptions (Auranen et al., 2002; Cantor et al., 2005; Coon et al., 2005; Shao et al., 2002; Trikalinos et al., 2006; Werling et al., 2014), largely failed to converge on replicable loci. Furthermore, linkage signals often implicate broad chromosomal regions containing many genes and thus are not informative at face value of the specific gene or genes harboring functional risk variants. Without implicating specific genes, linkage signals are therefore unable to inform genetic diagnoses or study of etiological mechanisms or potential treatments.

Other work identified connections between ASD symptoms and several different monogenic or single-locus syndromes, including Fragile X Syndrome, tuberous sclerosis, and others (Abrahams & Geschwind, 2008), which together suggested that ASD was more likely to have a multi-locus, complex genetic architecture. Most recently, models derived from the results of gene discovery studies of rare, *de novo*, dosage- or protein-disrupting variants in sporadic autistic cases have estimated this locus heterogeneity to be vast, with as many as 350-1000 genes involved in ASD risk (Iossifov et al., 2012; Sanders et al., 2012). As linkage signals are sensitive to the composition, or relative representation, of risk variants within the specific sample of

families used for each analysis, inconsistency in the loci reported by different linkage studies can reasonably be accounted for by heterogeneity of this degree.

Linkage signals, reported as logarithm of odds (LOD) scores, from studies of multi-family samples, represent an aggregation of signal across all families. Many linkage analyses for ASD have tested samples of mainly nuclear families where the affected members belong to a single generation; thus, instead of tracing the transmission of chromosomal segments through multiple generations of each family, LOD scores are derived from allele-sharing patterns between affected siblings. At any given autosomal locus, a sibling pair might share 0, 1, or 2 alleles, presumably inherited and identical by descent (IBD), and the null hypothesis assumes that 25% of sibling pairs will share 0, 50% will share 1, and 25% will share 2 alleles. LOD scores increase as the expected proportions of allele sharing across families shift toward a greater degree of sharing, and/or a lesser degree of not sharing, than expected. For this reason, testing different subsets of families carrying a wide range of risk variants will have an impact on the relative proportions of allele sharing at different genomic loci, leading to variable results across studies.

Another characteristic of linkage analysis is that the identity of specific alleles is only relevant within families, where siblings must share the same alleles for their family to contribute positively to the linkage signal. It is not necessary that different families share the same alleles as one another, only that their affected members share with each other at the same loci where other families also share. Therefore, in contrast with genetic association testing, which compares the frequencies of specific alleles between cases and controls or evaluates the transmission of specific alleles from parents to affected children, linkage signals are allele-agnostic. Furthermore, since alleles are inherited in haplotype blocks and signals are aggregated across

families, LOD scores often implicate large regions that include many genes. Given this coarse resolution, in order for a linkage signal to have utility for diagnosis, counseling, understanding pathophysiological mechanisms, or treatment development, the specific genes that are disrupted or altered within the broader linkage region must be identified.

4.2: Common variant association

Association testing is one approach for identifying a risk gene or allele by comparing allele frequencies for common single nucleotide polymorphisms (SNPs) between samples of cases and controls, or by evaluating deviations from expected allele transmission frequencies from parents to affected children. I have applied this second method, the transmission disequilibrium test (TDT), to the linkage regions we identified in AGRE families in chapter 3 on chromosomes 1, 6, 8, and 20, to determine whether particular genes within these broader regions are associated with ASD status. As described in chapter 3, we imputed additional SNPs to increase the density of coverage of these regions, and then I tested for association in the family subgroup in which the linkage signal was discovered (female, containing, FC, male-only, MO, or all families), and corrected all P-values for the number of independent SNPs within the region, as defined by a pairwise linkage disequilibrium $r^2<0.3$. Though there are SNPs within several regions that approach significance for association, none survive correction for multiple testing (Table 4.1).

Though I fail to observe significantly associated SNPs within these regions, the samples are too small to support a definitive conclusion that variants in these regions are not associated with ASD. Common variants associated with conditions that impact reproductive potential, including ASD, are predicted to have small effect sizes, as more disruptive variants are

Table 4.1: Top associated SNPs within each 2-LOD interval for suggestive linkage

Locus	Group	Left bound (bp, hg19)	Right bound	Total SNPs	Independ. SNPs	Top SNP	P-value	Adj. P-value
1p31.3	MO	56,961,756	67,002,765	20,907	1695	rs11207864	1.05e-03	1.778
4q13.1	ALL	42,652,004	81,152,247	83,652	5131	rs115667468	3.00e-05	0.154
4q26	MO	104,510,766	135,643,583	69,759	4296	rs17365679	5.26e-05	0.226
6q27	ALL	152,629,586	170,851,436	46,077	3794	rs4709139	9.28e-05	0.352
6q27	MO	156,377,582	170,851,436	34,340	2586	rs9355201	6.59e-05	0.170
8p21.2	FC	21,602,192	27,973,215	15,046	1240	rs78485638	4.21e-05	0.052
8p21.2	ALL	20,894,128	29,272,247	19,764	1749	rs2430804	5.11e-04	0.894
8p12	FC	27,973,215	48,515,523	30,499	1955	rs73228672	4.75e-05	0.093
8q13.2	ALL	64,104,092	82,503,999	40,188	3112	rs9643690	4.86e-05	0.151
20p13	ALL	63,244	4,817,968	11,022	1243	rs2273958	4.76e-04	0.591

Locus	Group	N transmitted	N un-transmitted	Odds ratio	Minor/major allele	Minor allele frequency	Gene	Location
1p31.3	MO	522	413	1.264	T/C	0.334	INADL	intron
4q13.1	ALL	29	70	0.414	G/A	0.012	NPFFR2	intron
4q26	MO	129	72	1.792	G/A	0.050	MAD2L1	~200kb downstream
6q27	ALL	47	20	2.350	T/G	0.011	RPS6KA2	~20kb upstream
6q27	MO	518	394	1.315	A/C	0.325	DACT2	~9kb upstream
8p21.2	FC	37	82	0.451	T/C	0.043	LOXL2	intron
8p21.2	ALL	444	555	0.800	G/C	0.047	RHOBTB2	exon
8p12	FC	22	57	0.386	T/C	0.032	PURG	~6kb downstream
8q13.2	ALL	929	758	1.226	T/C	0.287	FLJ39080	intron

under selection pressure and do not become common in the population. Therefore, very large samples are needed to implicate common variants with confidence. The set of all multiplex families tested for association here included a relatively large sample for ASD studies of 1,521 families, but this is still far smaller than samples used in many of the successful association studies for other complex psychiatric diseases, which utilize several thousand to tens of thousands of subjects (Ripke et al., 2011; Ripke et al., 2013; Sklar et al., 2011). Additionally, the intention behind stratifying these ASD families into subgroups of 606 families with only male children and 407 families with at least one female child was to assemble more genetically homogeneous samples, thereby increasing power to identify risk loci, though the smaller sample size of subgroups may offset the advantage provided by stratification. However, these family sets are an extension of the same samples and subgroups used in the linkage analysis to identify these regions, and therefore include the same families in which genetic signal was observed. By

imputing for variants not genotyped, I have tested for SNP association to the extent that is possible in this sample and find no robust evidence supporting a role for common variants in these linkage regions.

4.3: Rare variant association

Given that linkage analyses do not implicate specific alleles, but regions shared within and between families, it is possible that these shared regions may harbor any number of different variants, including those that are rare or even private to individual families. For example, risk variants from different families may occur at different base pair positions but may all impact the same gene within the region. Or, with the large number of risk genes predicted by *de novo* variant discovery studies, rare variants may even hit more than one gene within the same linkage region. If such risk variants are very rare or private, it is also unlikely that they would be in linkage disequilibrium with and tagged by a common SNP that was genotyped and tested for association. Therefore, genetic sequencing must be used to find these variants.

To identify rare variants within linked loci, one potential approach is to fully sequence all linkage regions, though to reduce sequencing costs and facilitate the interpretation of implicated variants, it may be preferable to target only coding and/or putatively functional sequences within the regions of interest. Applying this tactic for the linkage regions identified in AGRE families in chapter 3, I have selected target regions for a customized sequence capture protocol with a 7 Mb total target that includes gene exons, 1 kb upstream sequence from each gene's transcription start site to capture regulatory regions, long non-coding RNAs, other mRNAs and expressed sequence tags observed in human central nervous system tissue, DNase hypersensitive sites, transcription

factor binding site regions, regions with H3K4me3 histone marks, and mammalian conserved elements.

To further minimize costs, I also propose sequencing these targeted regions in a single affected subject from each family, under the assumption that affected siblings share the same risk variants within linked regions. For analysis of identified variants, I propose a modified case-control design based on each family's contribution to the LOD scores in each region of interest. Within each linkage region, individual families contribute positive, negative, or negligible signal to a LOD score, and these three classes correspond closely to the degree of allele sharing between affected children in the family, with either two, zero, or one allele in common. For a single region, I can then assume that families carrying the risk variants that are flagged by the LOD score peak share two or one allele IBD, and also that the risk variants carried by families with 0 shared alleles in the region are not likely to be at this locus. Therefore, on a region-by-region basis, affected children can be classified as cases if they share alleles with their affected siblings while other affected children can be classified as controls if they share zero alleles with their affected siblings (Table 4.2).

Table 4.2: Number of case and control subjects in each linkage region

Locus	1p31.3	6q27	8p21.2	8p12	20p13
Group	MO	ALL	FC	FC	ALL
1st tier cases, IBD=2	193	349	142	147	362
2nd tier cases, IBD=1	584	344	673	663	343
Controls, IBD=0	95	179	56	61	167
Cases/Controls	2.03	1.95	2.54	2.41	2.17

As a first-pass analysis, I will compare rare variants between cases with two shared alleles and controls with zero shared alleles. Then, since it has been suggested that variants driving ASD risk in multiplex families may follow a dominant inheritance pattern (Ronemus et al., 2014; Zhao et al., 2007) and would therefore only be present on one, shared chromosomal

segment, I will include the families who share one allele in the regional case sets for additional analysis. I will also prioritize variants for testing that are the most readily interpretable, starting with protein-disrupting variants (nonsense, spice site, frameshift, and non-synonymous), then adding missense variants, then variants in conserved regions, and finally including other intergenic, putatively functional regions. For tests of variants within gene transcripts, I will run a gene-based, pseudo-bi-allelic association test, in which each gene has two possible "alleles": one, where the gene contains a non-synonymous, protein-disrupting variant, and two, where the gene sequence is unaltered, or only harbors silent variants.

I have estimated my expected power to implicate rare variants with these methods in the linkage regions of interest from chapter 3 using a publicly available genetic power calculator (Purcell, Cherny, & Sham, 2003). For power calculations, since risk alleles will be observed directly via sequencing, I estimated both risk allele and marker allele frequency as the predicted per-gene frequency of non-synonymous variation given several different mutation rates. For each region, I calculate power with the per-generation, per-base, non-synonymous *de novo* mutation rate observed in sporadic autism cases (1.5e-08) (O'Roak et al., 2012; Sanders et al., 2012), as well as with higher variant rates, since some proportion of rare variants are also likely to be inherited in this multiplex family sample and will therefore be more common.

Estimated relative risk is dependent on the mutation rate, and is calculated here as the ratio of A) the proportion of cases carrying a non-synonymous variant in a gene, to B) the estimated per-gene variant frequency. I also assumed population prevalence for ASD to be 1%, which is an approximation of the ASD prevalence across the time frame in which the AGRE subjects were recruited. I considered all subjects from families sharing zero alleles IBD across the linkage region to be controls, and I combined adjacent FC linked regions at chromosome

8p21.2 and 8p12 into a single region referred to as 8p. For a case-control analysis for discrete traits, the predicted power for observing significant gene-based association within each region of interest is shown in Table 4.3.

I find that large discrepancies between the estimated per-gene allele frequency and the proportion of cases carrying a non-synonymous variant in a given gene result in extremely large relative risk estimates for which power cannot be reliably calculated. Thus, for low mutation rates, I identify and note the maximum number and rate of variant carriers among cases at which power can be calculated. For higher mutation rates, I note the *minimum* number and rate of variant carriers among cases at which 80% power is achieved. Last, given the number of cases with regional IBD=2 above the expected 25% of cases, I note the maximum mutation rate at which power above 80% is calculated. Increased mutation rate above this value leads to higher estimates of per-gene allele frequency and lower estimated relative risk. Therefore if I assume that the number of subjects from families with IBD=2 above expectation are the variant-carrying subjects, these analyses are well powered, but only if the identified variants are sufficiently rare. Currently, the rates of rare variants and the contribution of these rare variants, variants private to individual families, and *de novo* variants to the familial component of ASD risk are not known. This targeted sequencing study will provide one of the first looks at the characteristics of rare variants in high-risk families and will detail their contribution to the genetic architecture for multiplex ASD.

I note that power will be impacted if risk variants do not converge on the coding sequences of one or two genes at these loci. Additionally, this analysis paradigm will need to be adjusted if the putatively disruptive variants are located in targeted non-coding regions. To analyze these variants, I will need to either assign each non-coding element to a corresponding

Table 4.3: Estimated power to detect association for ASD status in families showing regional linkage with rare, non-synonymous variants in nearby genes

Locus, N cases IBD=2	Non-synonymous per-generation mutation rate (MR)	N cases carrying variants in one gene	Per-gene variant frequency in cases (N var. carriers / N cases)	Expected variant frequency (Mutation rate by ave. coding BP per gene in region)	Relative risk (Probability of variant within region given linkage vs. no linkage)	Power
1p31.3, 193 cases	1.50e-08	2	0.0104	1.1e-04	94.9090	0.2873[1]
	5.00e-08	7	0.0363	3.64e-04	99.6544	0.7221[1]
	1.00e-07	9	0.0466	7.28e-04	64.0635	0.8183[2]
	5.00e-07	11	0.0570	0.0036	15.6600	0.8380[2]
	7.00e-06	29*	0.1503	0.0510	2.9490	0.8094
6q27, 349 cases	1.50e-08	3	0.0086	1.09e-04	79.1217	0.4165[1]
	5.00e-08	9	0.0258	3.62e-04	71.2095	0.8418[2]
	1.00e-07	9	0.0258	7.24e-04	35.6048	0.8284[2]
	5.00e-07	11	0.0315	0.0036	8.7034	0.8135[2]
	1.55e-05	75*	0.2149	0.1123	1.9142	0.8252
8p, 147 cases	1.50e-08	1	0.0068	9.02e-05	75.4480	0.1473[1]
	5.00e-08	4	0.0272	3.01e-04	90.5376	0.4311[1]
	1.00e-07	10	0.0680	6.01e-04	113.1720	0.7885[1]
	5.00e-07	12	0.0816	0.0030	27.1613	0.8249[2]
	6.50e-06	24*	0.1633	0.0391	4.1787	0.8089
20p13, 362 cases	1.50e-08	2	0.0055	6.12e-05	90.2559	0.2675[1]
	5.00e-08	7	0.0193	2.04e-04	94.7687	0.6995[1]
	1.00e-07	9	0.0249	4.08e-04	60.9227	0.8045[2]
	5.00e-07	11	0.0304	0.0020	14.8922	0.8257[2]
	3.20e-05	88*	0.2431	0.1306	1.8615	0.8002

[1]Row shows the maximum estimable power given the mutation rate and corresponding relative risk.
[2]Row shows the number of variant observations required for 80% power
*Number of available subjects calculated as likely to be carrying a risk variant, given that they are in excess of the expected 25% of families who share two alleles IBD. In these rows, the highest mutation rate at which power remains above 80% is noted.

transcript, or to consider each variant-carrying target as a discrete element for pseudo-bi-allelic testing, as will be run for genes.

Variants of interest that I find to be associated will be confirmed by Sanger sequencing in the proband, and subsequently genotyped in all other family members in order to characterize the transmission and penetrance patterns of the variant. Specifically, I will confirm whether the variant is shared with the other affected sibling(s), as assumed. I will also investigate whether the variant is inherited or arises *de novo,* from which parent the variant originated or was transmitted, either by direct observation for inherited variants or by haplotype matching if *de*

novo, as well as whether unaffected male or female siblings in the family are carriers. Any genes or variants implicated by these methods may then be functionally analyzed to investigate the mechanisms by which the variants increase risk for ASD.

4.4: Potential implications

Linkage signals reported in the literature for ASD are highly inconsistent, implicating many different regions with little replication, and so it has been assumed that linkage analysis is an uninformative approach for investigating the genetic architecture of ASD. Certainly, studies of rare, *de novo* copy number variants and protein-disrupting single nucleotide variants in sporadic cases have proven more fruitful for discovering specific risk loci, but we still lack a basic understanding of the familial, heritable component of ASD risk – the very component that gave the field the first clues that ASDs were likely to be genetic conditions. The discovery of rare, deleterious variants within loci implicated by linkage signals would help to elucidate the mechanisms by which heritable variation increases ASD risk, for example, whether variants are predominantly inherited from a carrier mother, or if they result from germ line mosaicism, and so on. Such discoveries would also highlight a renewed relevance for linkage for the identification of genomic regions and families most likely to harbor functional, familial risk variants.

CHAPTER 5:

Gene expression implicates pathways at the interface between sexual dimorphisms and genetic risk variants for autism spectrum disorders

5.1: Abstract

Autism spectrum disorders (ASDs) are a group of pervasive developmental conditions with heterogeneous presentation and a comparably heterogeneous genetic architecture, with hundreds to 1000 risk genes expected as sample sizes continue to grow. ASDs also have a male-biased prevalence, and the mechanisms responsible for this sex difference in risk are not understood, although genetic findings support a role for female protective factors. On one hand, ASD risk genes could themselves show sexually dimorphic expression and function, or alternatively, they could interact with characteristic sexually dimorphic pathways. I reasoned that sex-differential gene expression patterns in healthy human neural tissue can be informative of sexually dimorphic processes that may overlap or interact with ASD risk genes, thus implicating potential mechanisms for male-biased risk. Here, I investigated sexually dimorphic gene expression levels in three independent data sets from adult and prenatal human neocortical tissue, and evaluated known ASD risk and associated gene sets for evidence of sex-biased expression. I find no evidence for systematic sex-differential expression of ASD risk genes. Instead, I observe that genes expressed at higher levels in males' than females' cortex are significantly enriched for astrocyte marker genes and genes up-regulated in ASD brain. These findings suggest that it is not sex-differential regulation of ASD risk genes, but rather naturally occurring sexually dimorphic processes, potentially including neuron-glial interactions, that modulate the impact of ASD risk variants in the human brain and underlie the sex-skewed prevalence of ASD.

5.2: Background

Autism spectrum disorders are a group of heterogeneous, pervasive developmental conditions characterized by deficits in social communication and restricted, repetitive behaviors

or interests (American Psychiatric Association, 2013) that are currently estimated to affect 1 in 68 children in the United States (Developmental Disabilities Monitoring Network, 2014). Genetic variation contributes strongly to ASD risk, as evidenced by high concordance rates for ASD among twins (Bailey et al., 1995; Hallmayer et al., 2011), high recurrence risk for siblings (Constantino et al., 2010; Ozonoff et al., 2011), overlap with monogenic syndromes such as Fragile X, Rett, and Timothy Syndrome (Abrahams & Geschwind, 2008), higher rates of large deleterious copy number variants (CNVs) (Levy et al., 2011; Pinto et al., 2010; Sanders et al., 2011; Sebat et al., 2007), and higher rates of rare *de novo* protein-disrupting single nucleotide variants (SNVs) in ASD cases compared with their siblings (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). While the set of genes that can now be definitively implicated as ASD risk genes has been growing rapidly, predictive models from the studies of SNVs in sporadic cases estimate that there are likely to be between 350 and 1000 genes involved in ASD risk (Iossifov et al., 2012; Sanders et al., 2012). Given the scope of this genetic heterogeneity, understanding the precise etiology of ASD and developing broadly applicable treatments has proven challenging.

One robust risk factor for ASD is sex: for every one female with ASD in the US, there are 4.5 affected males (Developmental Disabilities Monitoring Network, 2014), and this male bias in prevalence is consistent across countries and across diagnostic criteria (Fombonne, 2009). It is remarkable that ASD is perhaps the most notable of a number of neurodevelopmental disorders showing varying degrees of male sex bias, including attention deficit hyperactivity disorder (Polanczyk, de Lima, Horta, Biederman, & Rohde, 2007), specific language impairment (Tomblin et al., 1997), and Tourette syndrome (Freeman et al., 2000). Additionally, typically developing males and females show different trajectories in many facets of cognition and

behavior, including language and social cognition (Connellan, Baron-Cohen, Wheelwright, Batki, & Ahluwalia, 2000; Morisset, Barnard, & Booth, 1995), both of which are deficient in ASD. So, one model posits that ASD risk follows a multiple threshold liability model (Reich et al., 1975), with a higher minimum genetic liability required for females to manifest the ASD phenotype, given their relative social advantage (Tsai et al., 1981; Werling & Geschwind, 2013). This is also referred to as the female protective effect or female protective model (Ronald et al., 2013). One hypothesis that follows from the female protective model is that to be affected, females need to carry a greater genetic liability, or mutational load, than affected males. There is now evidence at the population-wide, family level, as well as at the genetic level to support this hypothesis, with siblings of female probands scoring higher on a quantitative measure of ASD traits than siblings of male probands (Robinson et al., 2013), and with a higher rate of deleterious CNVs and SNVs in female cases compared with males (Jacquemont et al., 2014). However, the FPE is not incompatible with the existence of male-specific risk factors, and the molecular mechanisms responsible for protecting females, or for increasing males' vulnerability to genetic risk, are currently unknown.

Several theories have been proposed regarding sex-differential risk and protective factors, including the concept of ASD as an X-linked disorder (Abrahams & Geschwind, 2008; Bailey et al., 1998; Gecz et al., 2009; Jamain et al., 2003), protective effects of imprinted, paternally expressed X chromosome genes (Skuse, 2000), and prenatal exposure to elevated levels of testosterone (Baron-Cohen, 2002). However, evidence supporting each of these theories is varied. Though several X-chromosome genes have been implicated in ASD, numerous autosomal genes contribute to risk as well, and protective, paternally expressed X-chromosome genes have yet to be identified. Genetic linkage studies also have suggested that autosomal loci,

rather than sex chromosomal loci, contribute to sex bias in ASD (Cantor et al., 2005; Schellenberg et al., 2006; Stone et al., 2004; Szatmari et al., 2007). Recent evidence has linked increased fetal testosterone levels to later autistic-like cognitive phenotypes (Auyeung et al., 2009; Auyeung et al., 2006; Chapman et al., 2006; Knickmeyer et al., 2005), as well as ASD diagnoses in a population sample (Baron-Cohen et al., 2014), but the details of the molecular and cellular mechanisms that translate this potential mechanism of prenatal hormone exposure to an ASD phenotype later in life are unknown.

I reasoned that since sex differences in gene expression patterns contribute to the organization and maintenance of a sexually dimorphic brain, evaluating genome-wide sex-differential gene expression in neural tissue can inform us of points of overlap with ASD risk genes and related pathways. Here, I test two basic hypotheses about the relationship between sexually dimorphic gene expression and ASD risk genes: (1) Autism risk genes are expressed at different levels in males and females. With sex-differential baseline expression levels, a deleterious or protein-disrupting mutation in a risk gene is likely to have effects of different magnitudes in males and females. Under this hypothesis, I expect to observe an enrichment of sex-differentially expressed (sex-DE) genes among known autism risk genes. (2) Autism risk genes are not expressed at different levels in males and females, but genes in interacting molecular pathways and/or cellular processes are differentially expressed by sex. Under this scenario, the downstream impact of sex-neutrally expressed ASD risk genes would be modulated by their gene products' interactions with processes that are sexually dimorphic. In this case, I expect to observe enrichment of sex-DE genes among gene sets representing processes associated with ASD pathophysiology, and not among known ASD risk genes themselves.

To evaluate these hypotheses, I analyzed RNA-sequencing (RNA-seq) and microarray gene expression data from the BrainSpan project and from an independent, in-house data set for evidence of sex-differential expression patterns (BrainSpan, 2013; Kang et al., 2011; Voineagu et al., 2011). To characterize stable sex differences in gene expression, I evaluated samples from adult subjects, and to characterize sex-differential expression during early development when ASD risk genes are highly expressed (Parikshak et al., 2013; Willsey et al., 2013), I also evaluated samples from prenatal subjects. I then tested multiple ASD risk and associated gene sets corresponding to my two hypotheses about the expression of ASD risk genes and associated pathways for enrichment of sex-differential expression (Basu, Kollu, & Banerjee-Basu, 2009; Cahoy et al., 2008; Darnell et al., 2011; Iossifov et al., 2012; Miller, Horvath, & Geschwind, 2010; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; Voineagu et al., 2011).

By characterizing sex-differential gene expression in human brain tissue, I find that beyond the Y-chromosome and the XIST transcript, sex differences in neocortical gene expression are subtle in amplitude. Still, in both adult data sets and in the prenatal neocortex, within the genes expressed at higher levels in males, I observe significant enrichments for genes belonging to the ASD-associated M16 co-expression module and for astrocyte markers. I find no significant enrichments of sex-DE genes expressed at higher levels in either males or females for any set of known ASD risk genes. These data are most consistent with my second hypothesis, and suggest that male-biased ASD risk may be related to sex-differential functioning of M16 module genes and/or sexual dimorphism in cortical astrocytes.

5.3: Materials and methods

5.3.a: Gene expression data from human brain tissue

Three gene expression data sets from post-mortem human brain tissue samples were analyzed for this study, including RNA-seq and exon array expression data from the BrainSpan project (BrainSpan, 2013; Kang et al., 2011), and RNA-seq data from an independent human cortical sample set produced at the University of California-Los Angeles.

5.3.a.i: Adult BrainSpan sample

BrainSpan RNA-seq data summarized to Gencode 10 (Harrow et al., 2006) gene-level reads per kilobase million mapped reads (RPKM) were used for the discovery stage. I opted to use the RNA-seq over the microarray data from the BrainSpan project because sex differences in gene expression beyond the Y chromosome and XIST in adult tissues are subtle (Yang et al., 2006) and RNA-sequencing allows for the detection of a wider dynamic range of gene expression levels which may aid my ability to detect meaningful sex differential gene expression patterns. These data were then normalized for GC content (Hansen, Irizarry, & Wu, 2012) and batch-corrected for processing site (Johnson, Li, & Rabinovic, 2007). Only samples from the frontal (dorsolateral prefrontal cortex, DFC; primary motor cortex, M1C; medial prefrontal cortex, MFC; orbitofrontal cortex, OFC; ventrolateral prefrontal cortex, VFC), temporal (primary auditory cortex, A1C; inferior temporal cortex, ITC; superior temporal cortex, STC), and parietal cortex (inferior parietal cortex, IPC; primary somatosensory cortex, S1C) from subjects aged 13-40 years with RNA integrity number (RIN) of at least 8.0 were used in this analysis.

RPKM values from samples meeting these criteria were then log-transformed ($\log_2$[RPKM+1]). Non-expressed genes and outlier samples were removed iteratively until all normalized inter-sample correlations fell within 2.5 standard deviations of the mean. Non-

expressed genes were defined as those genes with a log-transformed RPKM expression level of less than one in more than 50% of all male or female samples from the selected subset. Outlier samples were identified by evaluating inter-sample correlations and hierarchical clustering, first within each sex, and then on the full, non-stratified data set. To mitigate the effects of systematic differences in the range of expression levels across samples on the results of the differential expression analysis, I also performed quantile normalization. After gene filtering and outlier removal, 72 samples (29 from males, 43 from females) and 16,843 expressed genes remained. Given the wide discrepancy in the number of samples from male and female subjects, I then matched a subset of the female samples to the male samples on subject age and brain region and again filtered for expressed genes. The final data set consisted of 58 samples from 10 subjects (29 samples from 5 subjects of each sex) and 16,719 genes (Table 5.1).

Table 5.1: Discovery set adult subjects from BrainSpan

Subject	Sex	Age (years)	Ethnicity	Hemisphere	Site	pH	PMI (hours)	N cortex samples
HSB124	F	13	A	R	Yale	6.34	19.5	3
HSB119	M	15	A	L	USC	6.93	14.5	3
HSB105	M	18	E	L	USC	6.21	28	1
HSB127	F	19	E	L	Yale	5.91	9.5	3
HSB130	F	21	E	L	Yale	6.81	18	10
HSB136	M	23	A	R	USC	6.36	10.5	9
HSB126	F	30	E	R	Yale	6.92	9.5	8
HSB145	M	36	E	R	Yale	NA, imp. 6.52	18	9
HSB123	M	37	A	R	Yale	6.37	13	7
HSB135	F	40	A	R	USC	6.82	30.5	5

PMI, post-mortem interval. Imputed values (per-subject mean) noted by "imp."

5.3.a.ii: Adult replication sample

The adult replication set comprised 8 cortical samples from 5 male individuals and 8 cortical samples from 5 female individuals matched for age, post-mortem interval, and brain region. These samples were a subset of a larger set acquired from the Harvard Brain and Tissue Bank and the Eunice Kennedy Shriver National Institute for Child Health and Human

Development Brain and Tissue Bank for Developmental Disorders following the tissue acquisition policies of the respective brain banks. Two cortical regions were chosen from frozen brain samples: dorsolateral or medial prefrontal cortex (frontal cortex, from BA9), superior temporal gyrus (temporal cortex, from BA41, BA42, or BA22). Subject-level information from the brain banks included case/control status, age, sex, post-mortem interval, and medical history. Control status was confirmed by ensuring these individuals had no history of neuropsychiatric or neurological conditions.

Brain samples were dissected on dry ice in a dehydrated dissection chamber to reduce effects from sample thawing or humidity. Approximately 100mg of tissue across the cortical region of interest was isolated from each sample, and care was taken to keep samples at -80C to avoid RNA degradation. Up to two RNA isolations were performed for each sample using the miRNeasy kit (Qiagen). RNA quality was quantified using the RNA integrity number (Schroeder et al., 2006), with the same individual extracting all RNA.

Ribosomal RNA was depleted from 2ug total RNA with the Ribo-Zero Gold kit (Epicentre). Remaining RNA was then size selected with AMPure XP beads (Beckman Coulter) and resuspended, and subsequent steps followed the Illumina TruSeq protocol for library preparation with indexed adapters. Libraries were quantified with the Quant-iT PicoGreen assay (Life Technologies) and validated on an Agilent 2200 TapeStation system. Libraries were pooled to multiplex 24 samples per lane and each pool was sequenced six times on a HiSeq2000/2500 instrument using high output mode with standard chemistry and protocols for 50bp paired end reads. Raw read .fastq files were de-multiplexed using CASAVA (Illumina).

Reads were mapped to the human reference genome (hg19) using Gencode v18 annotations with TopHat2, allowing for up to 10 multiple mappings per read (Harrow et al.,

2006; Trapnell, Pachter, & Salzberg, 2009). Ouput .bam files were filtered to ensure that every read had a valid pair, resulting in only paired-end reads (fragments) being used for downstream analyses. Transcript levels were quantified using Gencode v18 gene models at the union gene model level using both HT-seq Counts (Anders, Pyl, & Huber, 2014) and Cufflinks. To filter for expressed genes, any given gene was required to have at least 10 counts in 80% of samples as determined by HT-seq Counts and a Cufflinks lower bound estimate of FPKM > 0 for 80% of samples. At this point, the sex of each sample was confirmed by evaluating the gene expression level of XIST and summed expression from chrY genes to ensure that males had a high chrY/XIST expression and females had high XIST/chrY expression (Hoen et al., 2013).

Using the genes called as expressed, high agreement between Cufflinks and HT-seq Count was observed (spearman's rho = 0.91), so HT-seq Counts FPKM values were used for analysis. These data were normalized for GC content biases using the cqn package in R (Hansen et al., 2012) which resulted in log2(Normalized FPKM) values, and ensured that there were no sample outliers with a summed sample correlation Z-score > 2 (Oldham, Langfelder, & Horvath, 2012). In the subset of control samples selected for the analyses presented here, filtering to genes expressed in the subset, outlier detection, and quantile normalization were performed as for the BrainSpan RNA-seq data, and the final filtered data set consisted of 16 samples from 10 subjects and 15,105 genes (Table 5.2).

5.3.a.iii: Prenatal BrainSpan sample

BrainSpan exon array data downloaded from the Gene Expression Omnibus (GSE25219) were used for the assessment of prenatal gene expression. Here, I opted to use the array over the RNA-seq data set from the BrainSpan project due to the greater number of samples available during prenatal stages, selected to begin after developing males' mid-gestation peak in

Table 5.2: Replication set adult subjects

Subject	Sex	Age (years)	PMI (hours)	Brain bank	Primary COD	Seq. batch	N cortex samples
UMB5168	F	16	NA (imp 21.77)	NICHD-BTB	Unknown	1	2
AN17425	M	16	26.16	Harvard-ATP	Heart attack	1	2
AN19760	M	28	23.25	Harvard-ATP	Unknown	1	1
AN15566	F	32	28.92	Harvard-ATP	Unknown	1	1
UMB5079	M	33	NA (imp 21.77)	NICHD-BTB	Unknown	1	2
AN08161	F	36	23.83	Harvard-ATP	Multisystem failure	1	2
AN10679	F	41	14	Harvard-ATP	Unknown	1	1
AN04479	M	44	23.26	Harvard-ATP	Unknown	1	1
AN15088	F	52	17.88	Harvard-ATP	Heart attack	2	2
AN11864	M	57	22.33	Harvard-ATP	Unknown	2	2

PMI, post-mortem interval. COD, cause of death. Imputed values (per-subject mean) noted by "imp."

Table 5.3: Prenatal subjects from BrainSpan

Subject	Sex	Age (PCW)	Ethnicity	Hemisphere	pH	PMI (hours)	N cortex samples
HSB154	M	16	AE	R	6.44	3	7
HSB96	M	16	H	R&L	NA (imp. 6.51)	2	8
HSB97	F	17	E	R&L	NA (imp. 6.51)	1	7
HSB100	F	19	A	R&L	6.56	4	8
HSB102	F	21	As	R&L	5.89	13	9
HSB99	F	21	As	R&L	NA (imp. 6.51)	2	13
HSB92	M	21	A	R	6.65	4	9
HSB159	M	22	E	R&L	6.58	2	13

PCW, post-conception weeks; PMI, post-mortem interval. Imputed values (per-subject mean) noted by "imp."

testosterone production (Niemi, Ikonen, & Hervonen, 1967; Smail, Reyes, Winter, & Faiman, 1981). Only samples from subjects between 16 and 22 post conception weeks (PCW) from the frontal, temporal, and parietal cortex and with RIN of 8.0 or greater were used in this stage of analysis.

Probe set IDs from the downloaded data were matched to Ensembl Gene IDs from Gencode 10 using the biomaRt function in R. Non-expressed genes were defined as those genes with a log-transformed median probe set intensity of less than 6 in more than 80% of all male or female samples within the selected prenatal subset, and were removed. Outlier samples were

detected as for the other data sets, and expression data were quantile normalized. After these processing steps, 134 samples (42 from males, 92 from females) and 10,014 expressed genes remained. To equalize the number of male and female samples in the data set, I then selected a subset of the female samples to match the male samples on age and brain region and again filtered for expressed genes. The final data set consisted of 74 samples from 8 subjects (37 samples from 4 subjects of each sex) and 9,865 genes (Table 5.3).

5.3.b: Differential expression analysis

Differential expression analyses for all data sets were performed using a linear mixed model and Bayesian t-tests as implemented in LIMMA (Smyth, 2005), a method that is particularly robust for analyzing small samples. For all analyses, sex was included as the main contrast in the regression model, subject was included as a random effect to correct for the non-independence of samples from the same individual brain, and covariates that showed significant correlations at $P<0.1$ with at least one of the first five principal components of the expression data were included in the model as fixed effects. Covariates included age, PMI, cortical lobe, and pH for the adult BrainSpan data, PMI for the replication data, and age, RIN, PMI, and lobe for the prenatal data. Per-subject average values were substituted for any missing PMI and pH values. Genes with a fold difference (FD) magnitude of at least 1.2 and an unadjusted P-value of 0.005, 0.01, or 0.05 were called as differentially expressed by sex (sex-DE).

5.3.c: Annotation gene sets

Sets of genes associated with ASD risk, expression patterns in autism brain, and neural cell types were selected for assessment of the sex-DE genes from all analysis stages.

5.3.c.i: Autism spectrum disorder risk genes

ASD risk gene sets included: 1) ASD candidate genes from the SFARI gene database, 2) genes with at least one rare *de novo* protein-disrupting SNV (RDNV) identified in sporadic ASD cases, 3) genes with at least one protein-disrupting or missense RDNV identified in ASD cases, and 4) gene targets of FMRP binding.

ASD candidate genes were selected from the SFARI Gene Autism Database, which catalogues autism candidate genes from published literature (Basu et al., 2009). To include only genes with the strongest genetic evidence for ASD risk, genes were filtered to only those classified as syndromic (category S) and evidence levels between 1-4. These criteria exclude genes with minimal evidence of association with ASD risk etiology, such as location within an ASD-associated CNV, near a GWAS SNP, within a linkage peak, overlapping a non-replicated association signal, or evidence of interaction with a high confidence risk gene. Of genes expressed in the full developmental time course of the BrainSpan RNA-sequencing data, there were 138 genes in this set of ASD candidates.

Next, since manually curated, literature-based candidate gene databases inherently favor those genes that have garnered more experimental investigation in the field, I also sought a gene set based on evidence from an unbiased, experimental screen. For this, intersecting sets of genes with RDNVs in sporadic ASD cases were compiled from four publications reporting the results of exome sequencing of the Simons Simplex Collection (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012). All genes with RDNVs in autistic probands from these four studies were compiled and classified as protein-disrupting (nonsense, splice site, or frameshift mutations), missense, or silent variants. Here, the set of genes with protein-disrupting

variants expressed in the BrainSpan data are used (116 genes), and the expanded set of expressed genes with either protein-disrupting or missense variants (598 genes).

Additionally, the set of transcripts that are binding targets of FMRP, the gene that is silenced in Fragile X Syndrome, are enriched for ASD risk genes (Iossifov et al., 2012) and so this functionally defined set was also evaluated for sex-differential expression. As the study to identify FMRP target genes was performed in mouse brain, here the 783 human orthologous FMRP target genes that are expressed in the BrainSpan data are tested.

5.3.c.ii: Gene expression patterns from ASD cortex

Gene sets showing autism-associated expression patterns were identified by a study of gene expression in post-mortem cortex from autistic adults (Voineagu et al., 2011) and included four sets: 1) genes expressed at significantly higher levels in ASD than control cortex, 2) genes expressed at significantly higher levels in control than ASD cortex, 3) autism-downregulated co-expression module M12, and 4) autism-upregulated co-expression module M16.

From the differential expression analysis performed by Voineagu and colleagues (Voineagu et al., 2011) across neocortex samples in the initial cohort, those genes with a fold difference (ASD vs. control) ≥ 1.2 and q-value≤ 0.1 were selected. These genes were split by fold difference direction to genes expressed at a higher level in ASD cortex (92 genes) and those expressed at a higher level in control cortex (145 genes). Two unsigned co-expression modules from the weighted gene co-expression network analysis run on these data were also selected, since these modules' eigengenes were significantly associated with ASD status. The M12 module showed general down-regulation in ASD samples, and is enriched for genes with neuronal and synaptic function, and for genes with evidence of sub-threshold common variant association signal; 419 members of the M12 module overlapped with genes expressed in the

BrainSpan data and are tested here. The M16 module showed general up-regulation in ASD samples and is enriched for inflammatory response and immune system functions. The M16 set consisted of 366 testable genes.

5.3.c.iii: Neural cell type markers

Since gene expression patterns in the brain have been shown to reflect the transcriptional activity of distinct neural cell types (Miller, Oldham, & Geschwind, 2008; Oldham et al., 2008; Winden et al., 2009), gene sets marking neurons, astrocytes, oligodendrocytes (Cahoy et al., 2008), and microglia (Miller et al., 2010) were culled from the literature. Within genes called as expressed across the full developmental time course of the BrainSpan RNA-seq, data, these sets included 1539 neuronal markers, 2015 astrocyte markers, 1647 oligodendrocyte markers, and 27 microglia markers.

5.3.c.iv: Background gene sets

For the following annotation analyses, the selection of an appropriate background gene set is necessary to draw valid conclusions about observed enrichments and depletions. For tests of the SFARI candidate genes, genes with RDNVs in ASD cases, and cell type markers, I used as background those genes with a Gencode biotype annotation of "protein-coding" and called as expressed in each data set. For tests of gene sets with ASD-associated expression patterns, I used as background those genes called as expressed and tested in the study by Voineagu and colleagues (Voineagu et al., 2011) and for tests of the set of FMRP interactors, I used the genome-wide set of one-to-one human-mouse orthologs.

For all background and gene sets of interest, gene identifiers were converted to Ensembl Gene IDs in Gencode using the biomaRt package in R to allow for unambiguous comparisons between genes from different data sources.

5.3.d: Over-representation analysis

Over-representation analysis was performed in order to annotate the sex-DE genes and to test two basic hypotheses about the interaction between sex differential gene expression and ASD risk genes and related pathways. Hypothesis 1 states that ASD risk genes are themselves differentially expressed by sex in human cortex. To test this hypothesis, I evaluated sex-DE genes for enrichment of the four ASD risk gene sets described above. Hypothesis 2 states that ASD risk genes are not differentially expressed by sex, but interact with pathways or processes that are sexually dimorphic in typical brain. To test this hypothesis, I evaluated sex-DE genes for enrichment of gene sets with ASD-associated expression patterns, and neural cell type markers also described above.

To evaluate the enrichment of these gene sets of interest among sex-DE genes, I applied a two-sided Fisher's exact test separately to sex-DE genes expressed more highly in males (male-up) and in females (female-up) at three P-value thresholds (unadjusted $P<0.005$, 0.01, and 0.05). Given the small sample sizes available in the tested data sets and subtle sex differences in gene expression levels beyond the Y chromosome, evaluating several sets of sex-DE genes at increasingly lenient P-value thresholds may allow for the observation of patterns of functional enrichments that are not apparent among the most stringently defined sex-DE genes. P-values from the Fisher's exact test were adjusted for the 24 tests performed against the 12 gene sets of interest at each significance level for the sex-DE input genes.

To further investigate underlying trends in the functional annotations of sex-differential expression patterns without applying arbitrary, hard cutoffs to define sex-DE genes, I also tested gene sets of interest for their mean gene rank within the differential expression results. In a modification of the metric utilized by Miller and colleagues (Miller, Woltjer, Goodenbour,

Horvath, & Geschwind, 2013), I first sorted and ranked all tested genes by their t-statistics from the differential expression test, such that significantly male-up genes had top ranks and significantly female-up genes ranked at the bottom. Ranks were then scaled from 1 to negative 1, with positive values corresponding to male-up genes and negative values corresponding to female-up genes. The scaled gene ranks for the members of each gene set of interest were then identified and averaged, and this mean gene rank was compared to the mean gene rank and distribution of the corresponding background gene set using a Z-test. All Z-test P-values were corrected for the 12 gene sets tested.

5.4: Results

5.4.a: Genes differentially expressed by sex in adult human cortex

The genome-wide pattern of sex-differential expression in adult cortical samples from BrainSpan shows expected, highly robust differential expression of Y chromosome genes and XIST, an X chromosome transcript responsible for initiating X chromosome inactivation that is only expressed in females (Figure 5.1A). I note that all Y-chromosomal transcripts that fail to show male-biased gene expression are pseudogenes with high sequence similarity to their corresponding genes on other chromosomes. Beyond the large expression contrasts for Y genes and XIST, the observed expression level differences between autosomal and non-XIST X chromosome transcripts are generally subtle in amplitude. For example, using a fold difference (FD) of 1.2 as a cutoff, I only identify 184 genes at $P<0.005$ (Figure 5.1B). I use multiple thresholds to capture patterns and trends in sex-differential gene expression to show that these results are not due to arbitrary statistical or FC cutoffs in a sample of this limited size (Figure

5.1). Results from enrichment analyses for gene sets of interest demonstrate the utility of this approach (see Figure 5.2).

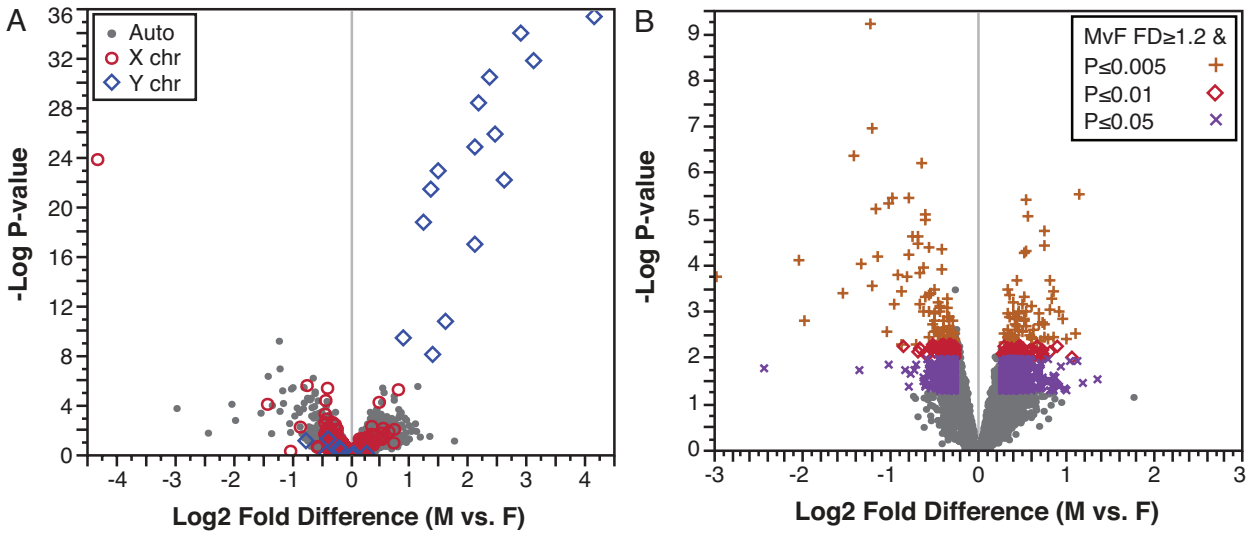


Figure 5.1: Sex-differential gene expression in adult human cortex from BrainSpan.
Sex differences in gene expression levels in human cortex are relatively subtle, with the exception of Y chromosome genes and XIST. A) Volcano plot for all 16,392 expressed transcripts, genes with positive log2 fold difference are expressed at higher levels in males, genes with negative log2 fold difference are expressed at higher levels in females. B) Subset of the volcano plot in A for the 15,723 autosomal transcripts.

5.4.b: ASD-associated pathways and cell types, but not ASD risk genes, show sex-differential expression in adult cortex

To test the hypotheses about the role of sexually dimorphic gene expression in ASD etiology, I first evaluated the identified sex-DE genes at three different P-value thresholds for enrichment of known ASD risk genes from several sources: a) candidate genes from a manually curated database (Basu et al., 2009), b) genes with rare, *de novo*, protein-disrupting or missense SNVs in sporadic ASD cases from the Simons Simplex Collection (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012), and c) FMRP binding targets (Darnell et al., 2011). Together, these gene sets capture a broad scope of ASD risk genes, from heavily studied

candidates, to genes with evidence of risk variation from exome-wide screening, to genes that participate in a regulatory network that is enriched for currently identified ASD susceptibility genes (Parikshak et al., 2013; Willsey et al., 2013). I found no evidence for significant enrichment of any tested set of known ASD risk genes, ranging from syndromic and candidate ASD-associated genes to genes identified by exome sequencing, providing no support for hypothesis 1 (Figure 5.2).

Next, I tested for sexually dimorphic expression in gene sets with evidence of ASD-associated expression patterns in neocortex, as well as gene sets corresponding to specific neural cell types (Cahoy et al., 2008; Miller et al., 2010). These ASD-associated gene sets come from a published study of gene expression in the post mortem cerebral cortex from autistic subjects (Voineagu et al., 2011), and include one gene co-expression module that was significantly down-regulated (M12) and another that was significantly up-regulated (M16) in ASD cortex. These differentially expressed genes and ASD-associated co-expression modules comprise large gene sets that are coherently altered in ASD cases with different genetic etiologies, and therefore likely represent either an upstream background of molecular risk for ASD, or the downstream consequences of deleterious variants in ASD risk genes, either of which might be sexually dimorphic.

In contrast to the lack of enrichment for ASD candidate risk genes, I observe several significant enrichments and depletions (fewer overlapping genes than expected) for cell type markers and gene sets with distinct expression patterns in ASD cortex, each of which becomes increasingly robust as the P-value threshold for inclusion is relaxed (Figure 5.2A). For genes expressed at significantly higher levels in males' cerebral cortex than females' cortex, I find a nearly five-fold enrichment of genes up-regulated in ASD cortex (ASD-up, 4.94-fold, adjusted

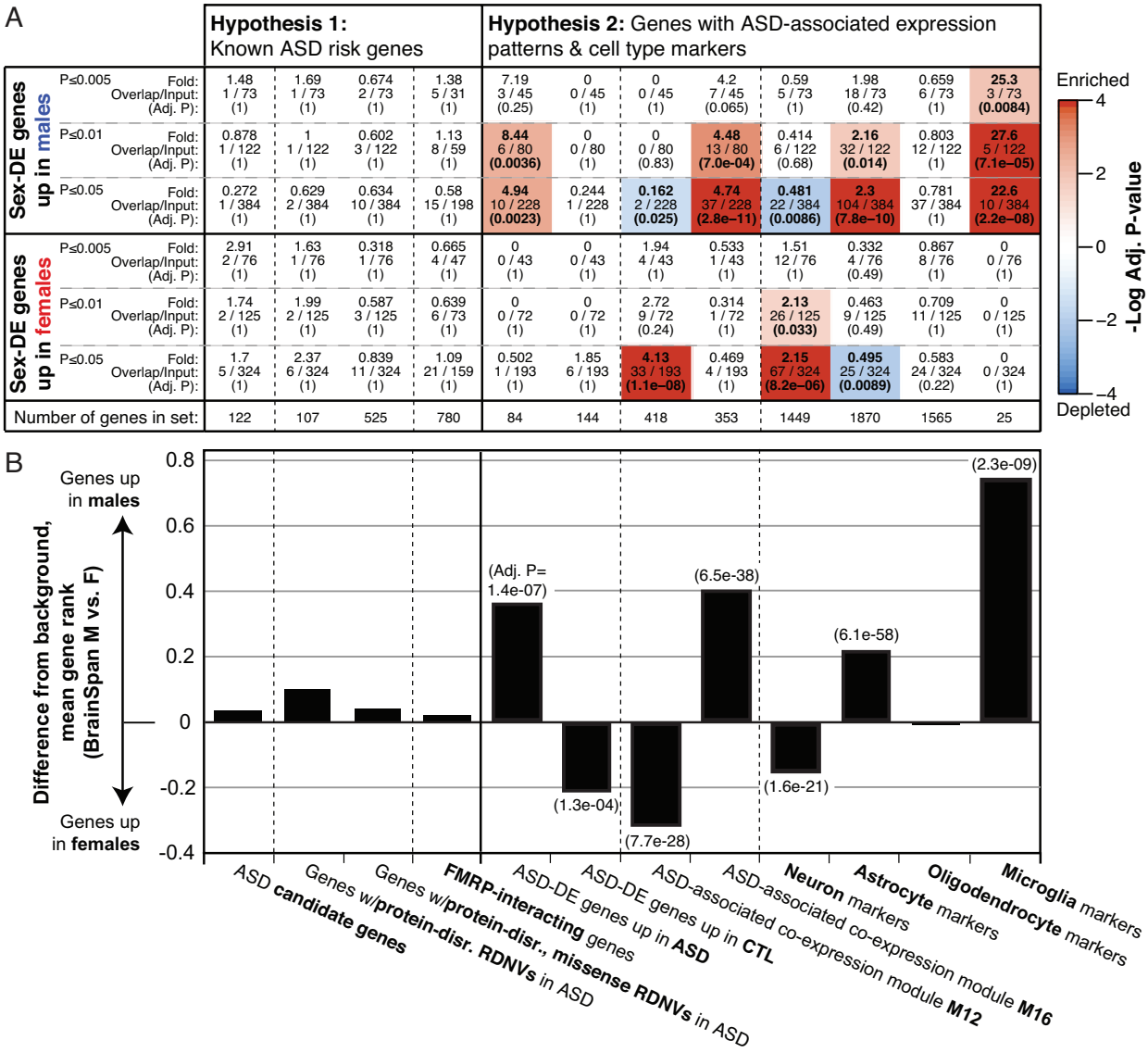


Figure 5.2: Sex-DE genes in adult human cortex are enriched for ASD-associated co-expression modules and cell type markers, but not for known ASD risk genes.
Over-representation analysis of sex-DE genes among known ASD risk genes (left) and gene sets with ASD-associated expression patterns and cell type-specific expression (right). A) Heat map showing results from Fisher's exact test on sex-DE genes defined by FD≥1.2 and several P-value thresholds. Enrichment P-values are adjusted for the 24 tests run at each threshold. All gene counts show the number of genes from each test and reference set that are present in the corresponding background gene set. B) Bar plot showing the difference between the mean gene rank of each test set and the corresponding background gene set. Mean gene rank represents the degree of expression direction skew of the members of gene set toward increasingly significant, differentially expressed genes and scales from -1 (perfect enrichment of test set genes for significantly female-up genes) to 1 (perfect enrichment of test set genes for significantly male-up genes), with 0 marking no difference from background in expression direction (see Materials and methods). Adjusted P-values (12 tests) from Z-tests against the mean and standard deviation of gene rank from the corresponding background set that are less than 0.1 are noted in the figure.

P=0.0023), and of genes belonging to ASD-associated module M16 (4.74-fold, adj. P=2.8e-11), both of which contain genes associated with astrocyte and microglial function. Consistent with this, I also observe a 2.3-fold enrichment for independently generated lists of astrocyte marker genes derived from purified mouse astrocytes (adj. P=7.8e-10) (Cahoy et al., 2008), and a 22.6-fold enrichment for microglia markers (adj. P=2.2e-08) (Miller et al., 2010). Parallel with the enrichment in M16, astrocyte, and microglial genes, genes belonging to the ASD-associated module M12 (0.162-fold, adj. P=0.025) and neuronal markers (0.481-fold, adj. P=0.0086) are significantly depleted among male-up DE genes. Genes expressed at significantly higher levels in females than males show fewer significant overlaps, but show reciprocal enrichments for those gene sets that are significantly depleted among male-up genes: the M12 module (4.13-fold, adj. P=1.1e-08), and neuronal markers (2.15-fold, adj. P=8.2e-06). Female-up genes are also depleted of astrocyte markers (0.495-fold, adj. P=0.0089), contrasting the enrichment observed for male-up genes.

To further evaluate the results of the sex-DE analysis for gene set enrichment trends without applying arbitrary thresholds, I then tested all gene sets of interest for significant shifts in the mean rank of the gene set members from background within the sex-DE results (Miller et al., 2013). Results from this more inclusive test corroborate findings from the Fisher's exact tests: ASD-up genes (adj. P=1.4e-07), M16 module members (adj. P=6.5e-38), astrocyte markers (adj. P=6.1e-58), and microglia markers (adj. P=2.3e-09) are expressed at relatively higher levels in males, while genes down-regulated in ASD cortex (CTL-up, adj.P=1.3e-04), M12 module members (adj. P=7.7e-28), and neuronal markers (adj. P=1.6e-21) are expressed at relatively higher levels in females (Figure 5.2B).

Since this study aimed to identify functional connections between sex-differential biology and the pathways and mechanisms involved in ASD pathophysiology, I was particularly interested in the significant enrichments of sex-DE genes in the ASD-associated co-expression modules, M16 and M12. Previous studies of co-expression patterns in brain tissue have demonstrated that co-expression modules correspond to neural cell types and coherent biological functions (Miller et al., 2008; Oldham et al., 2008; Winden et al., 2009); thus, sex-DE enrichments in these modules may most directly implicate those pathways involved in sex-biased ASD risk. I find that, of the 40 sex-DE genes that are also members of the M16 module, 39 genes are expressed at higher levels in ASD than control cortex, and 36 of these ASD-up, M16 genes are also expressed at higher levels in males than females (Figure 5.3A). I also observe that of the 35 sex-DE genes that belong to the M12 module, all 35 genes are expressed at higher levels in control than ASD cortex, and 33 of these CTL-up, M12 genes are also expressed at higher levels in females compared with males (Figure 5.3B). In short, I find concordant directionality of differential expression by sex and by ASD status, with M16 genes up-regulated in ASD and typical male brain and M12 genes up-regulated in control and typical female brain.

To assess whether these parallels in relative expression are restricted to the ASD-associated modules or if they reflect a more widespread gene transcription pattern common to the ASD and male brain (or to unaffected controls and the female brain), I compared the FDs from this sex-DE analysis with FDs observed in the differential expression analysis of ASD cortex (Voineagu et al., 2011). For the 418 sex-DE genes with FD≥1.2 and P≤0.05 also tested in the ASD-DE analysis, I observe a modest, but significant positive correlation between the sex-differential and ASD-differential FDs (r=0.30, P<1e-04; Figure 5.3C). This correlation improves to 0.45 (P<1e-04) for only those genes differentially expressed in both contrasts with FD≥1.2 and

P≤0.05 (N=181 genes). I note that among the 83 genes significantly up-regulated in both males and in ASD, 31 are M16 module members and 33 are astrocyte markers (21 are both astrocyte markers and M16 members); among the 68 genes significantly up-regulated in both females and in unaffected controls, 33 are M12 module members and 26 are neuron markers (14 are common to both M12 and neuron marker sets). These results suggest that the parallels in expression direction for genes by sex and by ASD status do appear to extend beyond the M16 and M12 modules, and that they may be more broadly indicative of relative transcriptional activity from specific neural cell types.

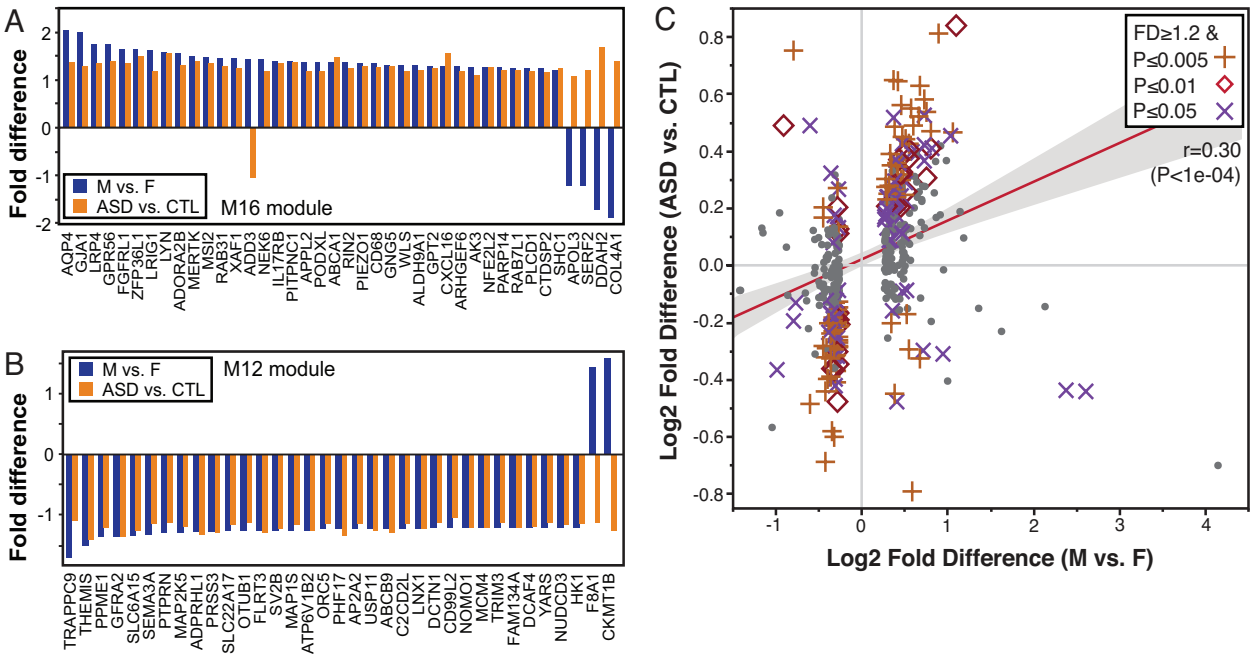


Figure 5.3: Sex-DE genes show distinct patterns related to ASD-associated pathways. Relative expression levels between males and females mirror expression patterns observed in ASD compared with controls from Voineagu et al. (2011). A) Sex-DE genes (FD≥1.2 and P<0.05) that also belong to the ASD-associated M16 co-expression module are mostly up-regulated in both male and ASD cortex as compared with females and controls. Genes with positive fold differences are expressed at higher levels in males and in ASD, and genes with negative fold differences are expressed at higher levels in females and in controls. B) Sex-DE genes that also belong to the ASD-associated M12 co-expression module are mostly down-regulated in both male and ASD cortex as compared with females and controls. C) Sex-DE genes' fold differences by sex (x-axis, N=418 genes with FD ≥1.2 and P <0.05) are positively correlated (r=0.30, P<1e-04) with the fold differences in ASD versus control cortex samples (y-axis). The best fit line is plotted in red with a 95% confidence interval shaded in gray. Colored marks note the significance level of ASD-DE genes with FD ≥1.2.

5.4.c: Independent replication of adult sex-differential gene expression

To validate these observations from the adult BrainSpan data, I then analyzed RNA-seq data from an independent sample of adult cortex tissue (15-57 years) for sex-differential expression patterns. This sample was of comparable effective size as the adult BrainSpan data, with data from 10 subjects (5 females), but included fewer regional samples from each individual with an overall total of 16 frontal and/or temporal cortex samples. Using this smaller data set, at a minimum FC magnitude of 1.2, I identify 53 genes at P<0.005, 95 genes at P<0.01, and 431 genes at P<0.05 (Figure 5.4B); only 19 genes are differentially expressed at a BH adjusted P-value of 0.05, all of which are on the Y chromosome. Testing for significant over-representations of ASD risk genes, expression patterns, and cell type markers among the sex-DE genes from this data set again failed to identify significant enrichments in any set of known ASD risk genes. In fact, the top 1000 sex-DE genes (FD≥1.2, ranked by P-value) from these results overlapped with significantly fewer FMRP interacting genes than expected among both the male-up genes (0.393-fold, adj. P=0.013) and among the female-up genes (0.273-fold, adj. P=0.0074). I do find significant enrichments for M16 module and astrocyte marker genes in the male-up genes, and these findings are echoed by the mean gene rank tests (M16 adj. P=4.1e-13; astrocytes adj. P=2.1e-41). In these data, I also find a significant depletion of oligodendrocyte markers among male-up genes (0.561-fold, adj. P=0.033). In direct contrast to the BrainSpan data, I see a significant enrichment for microglia markers among female-up genes (9.18-fold, adj. P=0.040), though I note that the tested set of microglial markers is quite small (N=25 genes) and is therefore potentially sensitive to differences between data sets. Overall, there was a high agreement between the expression levels of sex-DE genes from the BrainSpan data in this independent sample (r=0.83, P<1e-04). Also, of the 84 genes expressed more highly in males

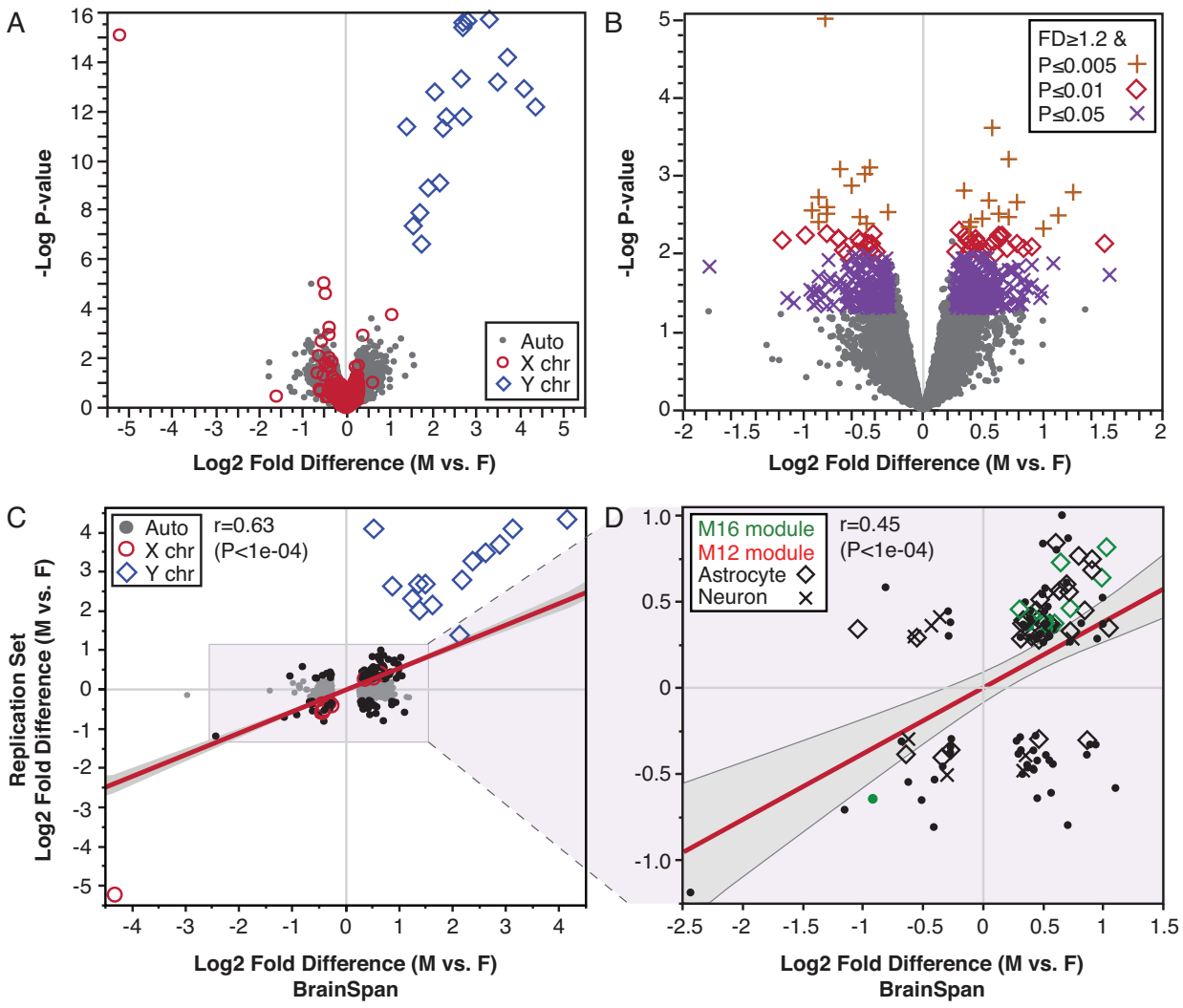


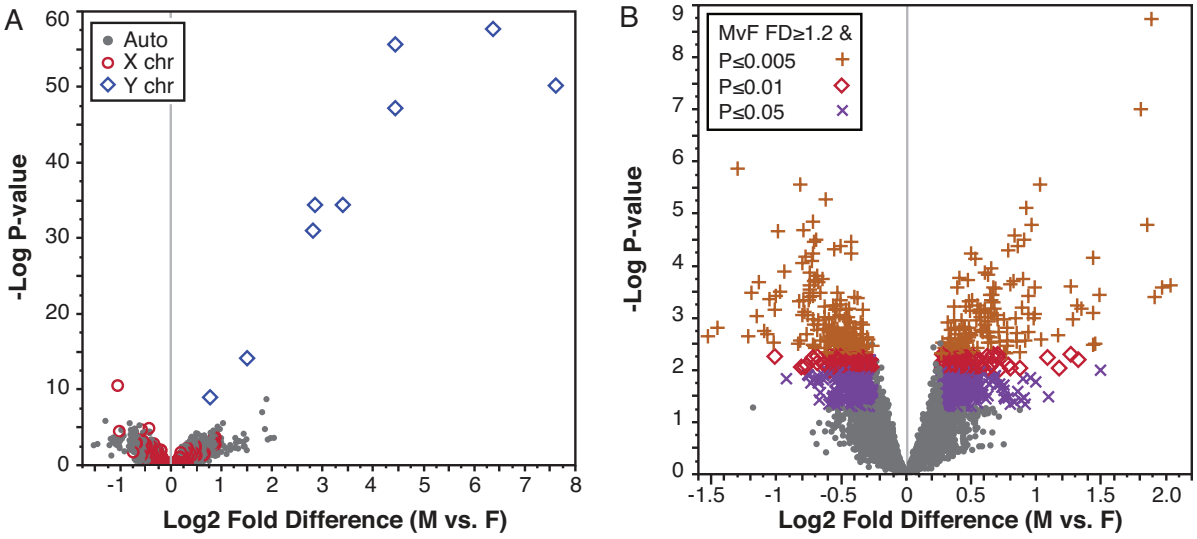
Figure 5.4: Independent replication of sex-differential gene expression.
An independent adult cortex sample shows parallel patterns of sex-differential gene expression compared with the adult BrainSpan data. A) In the replication sample, Y chromosome genes and XIST show robust sex differences in gene expression levels; volcano plot for all 15,105 transcripts expressed in the replication sample. B) Subset of the volcano plot in A for the 14,565 autosomal transcripts expressed in the replication sample. C) Fold differences for sex-DE genes from the BrainSpan sample also tested in the replication set (x-axis, N=672 genes with FD ≥1.2 and P <0.05) are positively correlated with fold differences observed in the replication sample (y-axis, r=0.63, P <0.0001). Best fit line is plotted in red with a 95% confidence interval shaded in gray; gray points note genes with FD <1.2 in replication sample. D) Autosomal sex-DE genes from BrainSpan with FD ≥1.2 in the replication sample (N=121 genes).

from both data sets with a FD≥1.2, 13 are M16 module members (9.25-fold enrichment, adj. P=7.1e-07) and 32 are astrocyte markers (4.26-fold enrichment, adj. P=1.6e-07), further

illustrating the robustness of male-biased expression levels of the ASD-associated M16 module and astrocytic genes in adult human cortex (Figure 5.4D).

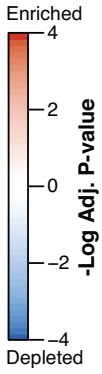
5.4.d: Genes differentially expressed by sex in prenatal human cortex also show enrichment for ASD-associated pathways and cell types

Recent investigations of the spatiotemporal expression profiles of ASD risk genes have implicated the developing fetal neocortex as a key developmental window where ASD risk genes are highly expressed (Parikshak et al., 2013; Willsey et al., 2013). To address the issue that I fail to detect significant enrichments of ASD risk genes among sex-DE genes because I am examining the wrong stage of development, I next analyzed microarray expression data from the BrainSpan project generated from 74 neocortex samples from 8 subjects (4 females) between 16-22 PCW. This prenatal epoch follows the mid-gestation peak in testosterone secretion from males' differentiated testes (Niemi et al., 1967; Smail et al., 1981). I use microarray data here instead of RNA-seq to increase power because there are a greater number of samples characterized by microarrays. At a minimum FD of 1.2, I find 330 sex-DE genes at $P<0.005$, 502 sex-DE genes at $P<0.01$, and 1,118 sex-DE genes at $P<0.05$, with 84 genes sex-DE at an adjusted $P<0.05$ (Figure 5.5A-B). Testing for enrichments of the gene sets of interest within these prenatally sex-DE genes reveals patterns that mirror those in the adult data: there are no significant enrichments for ASD risk gene sets within either male-up nor female-up genes, and FMRP interactors are significantly depleted within both male-up and female-up genes (Figure 5.5C-D). Genes expressed more highly in prenatal male cortex also show robust enrichments for ASD-up genes (5.19-fold, adj. P=1.1e-04), M16 module genes (3.24-fold, adj. P=6.5e-07), and astrocyte markers (1.61-fold, adj. P=9.8e-04), and a significant depletion of neuron marker genes



C

		Hypothesis 1: Known ASD risk genes				Hypothesis 2: Genes with ASD-associated expression patterns & cell type markers							
Sex-DE genes up in males	P≤0.005 Fold:	2.66	1.53	1.58	0.32	5.71	0.58	0.543	**3.91**	**0.233**	**2.29**	0.438	0
	Overlap/Input:	4 / 141	2 / 141	10 / 141	4 / 77	5 / 104	1 / 104	3 / 104	13 / 104	5 / 141	41 / 141	9 / 141	0 / 141
	(Adj. P):	(1)	(1)	(1)	(0.34)	(0.067)	(1)	(1)	**(2.3e-03)**	**(3.5e-03)**	**(7.0e-04)**	(0.29)	(1)
	P≤0.01 Fold:	2.16	0.982	1.65	**0.263**	**5.42**	0.38	0.353	**4.09**	**0.21**	**1.85**	0.764	0
	Overlap/Input:	5 / 217	2 / 217	16 / 217	5 / 115	7 / 157	1 / 157	3 / 157	20 / 157	7 / 217	54 / 217	23 / 217	0 / 217
	(Adj. P):	(1)	(1)	(1)	**(0.017)**	**(0.015)**	(1)	(1)	**(2.3e–05)**	**(7.8e–06)**	**(6.3e-03)**	(1)	(1)
	P≤0.05 Fold:	1.29	0.629	1.44	**0.219**	**5.19**	0.489	0.674	**3.24**	**0.472**	**1.61**	0.867	5.32
	Overlap/Input:	7 / 497	3 / 497	32 / 497	11 / 292	14 / 362	3 / 362	13 / 362	36 / 362	35 / 497	110 / 497	59 / 497	3 / 497
	(Adj. P):	(1)	(1)	(1)	**(3.8e–08)**	**(1.1e-04)**	(1)	(1)	**(6.5e–07)**	**(1.3e-04)**	**(9.8e-04)**	(1)	(0.70)
Sex-DE genes up in females	P≤0.005 Fold:	1.09	1.3	0.769	0.374	0	0.523	0.322	0.228	0.833	0.799	1.16	0
	Overlap/Input:	2 / 165	2 / 165	6 / 165	5 / 83	0 / 115	1 / 115	2 / 115	1 / 115	19 / 165	21 / 165	25 / 165	0 / 165
	(Adj. P):	(1)	(1)	(1)	(0.63)	(1)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
	P≤0.01 Fold:	1.45	1.73	0.75	0.344	0	0.338	0.314	0.147	0.894	0.734	1.26	0
	Overlap/Input:	4 / 253	4 / 253	9 / 253	7 / 125	0 / 176	1 / 176	3 / 176	1 / 176	31 / 253	30 / 253	41 / 253	0 / 253
	(Adj. P):	(1)	(1)	(1)	(0.066)	(1)	(1)	(0.86)	(0.54)	(1)	(1)	(1)	(1)
	P≤0.05 Fold:	1.54	0.973	0.979	**0.502**	0.56	1.55	0.55	0.493	0.843	0.663	0.867	0
	Overlap/Input:	9 / 547	5 / 547	25 / 547	22 / 273	2 / 370	9 / 370	11 / 370	7 / 370	64 / 547	60 / 547	65 / 547	0 / 547
	(Adj. P):	(1)	(1)	(1)	**(0.033)**	(1)	(1)	(1)	(1)	(1)	(0.065)	(1)	(1)
Number of genes in set:		103	87	432	700	62	108	340	242	1250	1428	1240	13



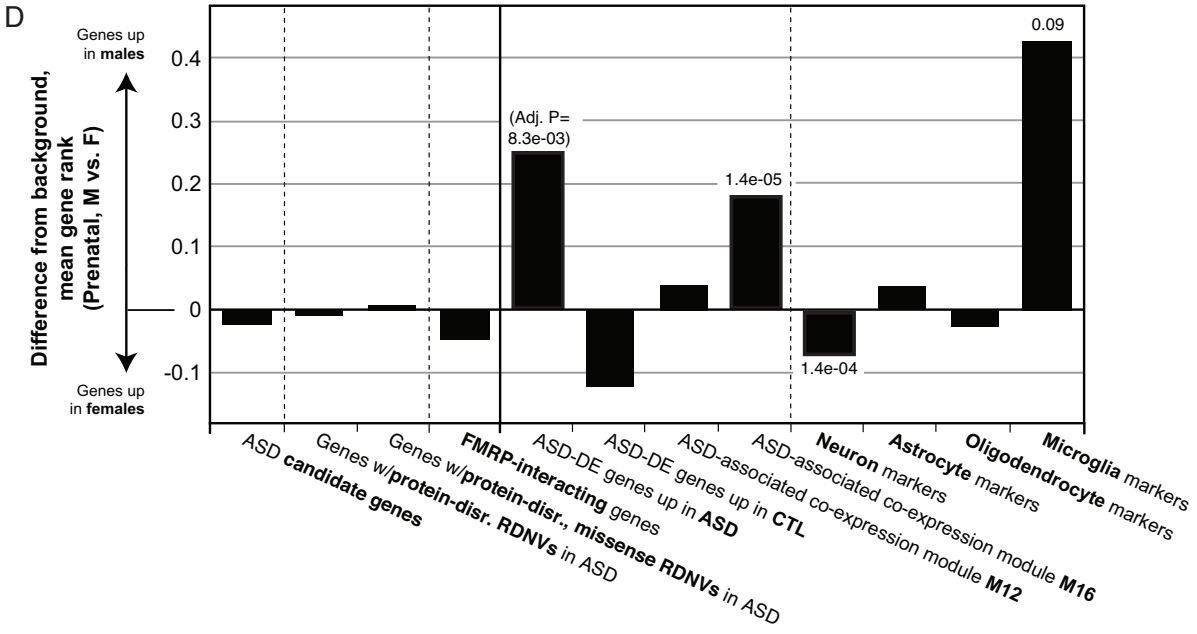


Figure 5.5, see caption on next page.

Figure 5.5: Sex-differential gene expression in prenatal human cortex from BrainSpan. Genes show sex-differential expression in the prenatal cortex, and sex-DE genes show robust enrichments and depletions for gene sets of interest. A) Sex differences in gene expression levels in prenatal human cortex are robust, with Y chromosome genes showing the most pronounced differential expression (XIST not tested). Volcano plot for all 9820 expressed transcripts. B) Subset of volcano plot in A for the 9467 autosomal transcripts. C) and D) Over-representation analysis of sex-DE genes among known ASD risk genes (left) and gene sets with ASD-associated expression patterns and cell type-specific expression (right). C) Heat map shows the results from Fisher's exact test on the prenatally sex-DE genes at several P-value thresholds with FD≥1.2. Enrichment P-values are adjusted for the 24 tests run at each threshold. All gene counts show the number of genes from each test and reference list that are present in the corresponding background gene set. D) Bar plot showing the difference between the mean gene rank of each test set and the corresponding background gene set; positive scores indicate a shift from background toward significantly male-up genes and negative scores indicate a shift from background toward significantly female-up genes. Adjusted P-values (12 tests) from Z-tests against the mean and standard deviation of gene rank from the corresponding background set that are less than 0.1 are noted in the figure.

(0.472-fold, adj. P=1.3e-04). (Figure 5.5C-D). Together with the analyses of gene expression in adult cortex, these findings do not support the hypothesis that ASD risk genes are differentially expressed by sex. These observations are instead more consistent with the hypothesis that ASD risk is modulated by sex due to the interactions of ASD risk genes with sexually dimorphic cellular and molecular processes, which likely include the behavior of astrocytes and functions of genes belonging to the ASD-associated M16 module.

5.5: Discussion

I characterized sex-differential gene expression levels in male and female neocortex samples from the adult and prenatal human brain and used these profiles to explore potential mechanisms by which sexually dimorphic biology modulates ASD risk. I find that genes implicated in ASD risk are not enriched for sex-differential expression in the adult or prenatal cortex. Rather, I observe that genes up-regulated in post mortem ASD brain, largely corresponding to genes highly expressed in astrocytes or other glial cells, are significantly enriched among genes expressed at higher levels in the adult and prenatal male cortex. Beyond

defined gene sets of interest, I also observe that the relative gene expression levels in males compared with females parallels that observed in ASD cases versus controls, genome-wide: male-up genes are also expressed at higher levels in ASD cortex and female-up genes are also expressed at higher levels in controls. This pattern in ASD versus controls is not likely due to sex bias, as in the referenced analysis of differential gene expression in autistic and control cortex, female subjects comprised a greater proportion of the cases (4 of 16) than the controls (1 of 16), and neither ASD-associated module shows a significant association with sex (Voineagu et al., 2011).

Previous studies of sex differences in gene expression in human brain have preferentially reported those sex-DE genes that show pronounced differential expression levels (Kang et al., 2011; Trabzuni et al., 2013). As a result, the set of genes known to be sex-DE in the human brain is heavily comprised of Y and X chromosome genes and does not include more subtle sex-differential expression, which may be more indicative of sex-specific tuning of molecular pathways than of more prominent sex chromosome copy number differences. Here, to capture sex-differential expression of genes that participate in molecular and cellular processes related to ASD, I applied a range of significance thresholds for calling genes as differentially expressed. I note that the FDs and P-values for calling differentially expressed genes are inherently arbitrary, and emphasize that despite the almost certain inclusion of false positive sex-DE genes in over-representation analyses, I observe enrichments for functional gene sets that replicate across independent samples. These enrichments are corroborated by observed shifts in the mean gene rank metric, which I used to assess the genome-wide distribution of differential expression for coherent skewing of interesting gene sets toward biased expression in one sex or the other (Miller et al., 2013).

I used sex-DE genes to test for enrichment within gene sets corresponding to two hypotheses that address the potential connections between sex-specific factors and genetic risk variants. The first supposes that genes harboring ASD risk variants are differentially expressed by sex and thus lead to different impact in males and females. While I do find that a few ASD candidate genes, RDNV-carrying genes, and FMRP interactors meet criteria for differential expression, sex-DE genes are not significantly enriched within these gene sets in either adult data set, or in the prenatal cortex. Notably, in the adult replication and prenatal data sets, I find significant depletions of FMRP interactors within both the male-up and female-up sets of sex-DE genes. These results suggest that ASD risk genes on the whole, and the FMRP regulatory pathway in particular, comprise fundamental functions in human neural development and function that are not directly regulated by sex-differential factors. It is still possible that the tested risk genes are differentially regulated in brain regions other than the neocortex, or at other stages of development, though I note that I observe these sex-neutral expression patterns in the mid-fetal neocortex, the spatiotemporal point of convergence for expression of ASD risk genes (Parikshak et al., 2013; Willsey et al., 2013).

The second hypothesis, which is supported by my analysis, posits that genes from downstream or interacting pathways, but not ASD risk genes themselves, will be differentially expressed by sex. Results from enrichment analyses of genes with ASD-associated expression patterns and neural cell type markers support this second hypothesis, showing consistent enrichments of the ASD-associated M16 co-expression module and astrocyte markers within male-up genes. The M16 module genes were defined in post mortem brain and are enriched for immune system function and inflammatory responses and are up-regulated in ASD brain (Voineagu et al., 2011). This is consistent with the genetic evidence presented in that paper,

where the M16 module is not enriched for ASD susceptibility variants and therefore is considered to represent a secondary, likely downstream effect of genetically causal perturbations (Voineagu et al. 2011). Whether a secondary effect of causal variants or a result of independent mechanisms, it may be that expression of this gene set beyond a certain threshold is detrimental to neural development and function. My findings show that typical males land closer than females to this putative threshold, potentially implicating the M16 module in mechanisms driving the male-bias in ASD risk. Furthermore, to the extent that M16 up-regulation contributes to the pathophysiology and manifestation of an autistic phenotype, this module may serve as a potential target for pharmacological treatments that could modulate the effects of numerous, heterogeneous risk variants acting upstream. I also find enrichments of astrocyte markers among male-up genes in all data sets tested, and there are two possible explanations for this: the first is that a greater number of astrocytes relative to neurons in male cortex than in female cortex could result in higher measured astrocytic gene expression in males, and the second is that males' cortical astrocytes may be more transcriptionally active than females'.

Aside from early, tenuous observations of greater cortical neuron density in males (Rabinowicz, Dean, Petetot, & de Courten-Myers, 1999), sex differences in the cellular composition of the human cortex have not been sufficiently characterized to determine if there are a greater number of astrocytes in male cortex. In contrast, there is abundant evidence for differences in astrocyte morphology between males and females, with male astrocytes possessing a greater number of longer and more branched processes, though the majority of these findings come from study of hypothalamic nuclei (Garcia-Segura, Lorenz, & DonCarlos, 2008; Mong, Kurzweil, Davis, Rocca, & McCarthy, 1996; Mong & McCarthy, 1999, 2002). Observations of male-typical astrocyte morphology in female animals exposed to testosterone neonatally (Mong

& McCarthy, 1999) and of estrogen receptor expression in hypothalamic and hippocampal astrocytes (Azcoitia, Sierra, & Garcia-Segura, 1999; Garcia-Segura et al., 2008; Sakuma, Tokuhara, Hattori, Matsuoka, & Yamano, 2009) are consistent with a role for sex steroid hormones in astrocyte differentiation and regulation. Given their role in modulation of neurotransmission (Araque, Parpura, Sanzgiri, & Haydon, 1999; Bezzi et al., 1998; Perea, Navarrete, & Araque, 2009) and synapse formation and function (Haber, Zhou, & Murai, 2006; Hama, Hara, Yamaguchi, & Miyawaki, 2004; Muller, 1992; Muller & Best, 1989), it is plausible that sexual dimorphisms in astrocyte number or function would have sex-differential effects on neuronal connectivity (McCarthy et al., 2003). It is also interesting to consider that numerous ASD risk genes function at the synapse (Auerbach, Osterweil, & Bear, 2011; Bourgeron, 2009; Gilman et al., 2011; Iossifov et al., 2012; Neale et al., 2012; Sakai et al., 2011), and that as an external, third party to the pre- and post-synaptic neurons, sexually dimorphic astrocytes are well positioned to broadly influence the effects of upstream, heterogeneous risk variants.

I caution against interpreting the enrichments that I observe among genes expressed at higher levels in males as male-specific or male-biased processes. This is because differential expression analyses for sex do not identify directional changes in expression, but relative differences in expression levels between the sexes. The true direction of these differences depends on the regulatory mechanism in place at the molecular level, for example, whether a gene is activated by upstream androgen receptor binding or whether it is repressed by estrogen receptor binding. Experimental manipulation of hormone exposure or transcription factor activity is needed to conclusively determine whether these genes are truly up-regulated in males, whether they are down-regulated in females, or both. Nevertheless, genes expressed more highly in typical male cortex also tend to be more highly expressed in ASD cortex, while genes more

highly expressed in typical female cortex also tend to be expressed at lower levels in ASD cortex (Voineagu et al., 2011). I note that this concordance is not likely the result of testing a male-skewed ASD sample, since this ASD expression study analyzed mostly male tissue, with 4 of 16 female ASD cases and 1 of 16 female controls. Interestingly, the observed gene expression patterns fit nicely with hypotheses derived from both the extreme male brain theory and from the concept of female protective factors in ASD: male-up/ASD-up genes could be acting as risk genes, while female-up/ASD-down genes could function as protective genes. Studies including sufficient numbers of both female and male cases and controls will be necessary to tease apart how sex and ASD status interact with regards to gene expression, for example, to determine whether male-up genes are expressed at even higher levels in ASD cases.

5.6: Conclusions

I find no evidence of sex-differential expression of ASD risk genes in the neocortex during the developmental epochs tested. Future studies in other brain regions, at other stages of development, and with a greater number of samples will be informative of the consistency of these patterns, which suggest that it is the molecular, cellular, or circuit-level context in which risk genes operate that is responsible for modulating ASD risk in a sex-differential manner. This notion fits with the remarkable spatiotemporal consistency of the male bias in ASD prevalence, despite ever-growing evidence of dramatic genetic heterogeneity in ASD. These analyses suggest that evaluation of the relationship and interactions between astrocytes and synaptic function and the causes and functional consequences of an up-regulated M16 module is needed to determine how sex-differential functioning of these pathways influence the neurodevelopment of a brain at risk for ASD.

CHAPTER 6:

Conclusions, limitations, and future directions

In this dissertation, I aimed to address the current gap in understanding of the biological mechanisms responsible for the male bias in ASD prevalence, a consistent feature of ASDs that is likely to be linked to its fundamental pathophysiological processes. Since its definition as a neurodevelopmental condition, at different times, under different versions of diagnostic criteria, and in different countries, a greater number of males have been diagnosed with ASDs than females, and this bias is especially pronounced among individuals with high adaptive functioning and IQ. However, epidemiological studies that apply unbiased community screens find more affected females than are observed in studies using existing diagnostic records, which suggests that females may be under-diagnosed and that the presentation of ASD in females differs from that represented by male-based diagnostic criteria. In either case, sex-specific biological factors are likely to play a modulatory role in the processes responsible for the sexually dimorphic risk for, or presentation of, ASDs. An understanding of these modulatory processes may also be informative of basic developmental mechanisms driving the presentation of the ASD phenotype at a molecular, cellular, or circuit level, though currently the specifics of how sex-differential factors operate with ASD's etiology are not known. In contrast, it is well known that genetic variation contributes substantially to ASD risk, and so in this work I have leveraged current knowledge about the genetics of ASD to investigate relevant sex-differential biology, with the long-term goal to identify and understand processes that put males at risk and/or protect females.

First, I have evaluated children from multiplex families and twin pairs from the commonly studied AGRE collection for recurrence and concordance rates consistent with a sex-differential threshold liability model, or female-protective model. A key prediction from this model is that affected females carry more deleterious variant loads in their genomes than the average affected male, and under the assumption in multiplex families that these variant loads

are mostly heritable, affected females share these more deleterious variants with their siblings, which increases siblings' risk. This has implications for genetic counseling as well as genetic study design, since variant discovery studies focused on female cases may be more likely to identify the large effect size variants carried by females and their siblings.

Here, to facilitate study of sex-differential, familial risk variants in multiplex families, I evaluated recurrence and concordance rates in families from the Autism Genetics Resource Exchange (AGRE) and find higher recurrence to male children, as well as to siblings and co-twins born into families or twin pairs with at least one affected female (FC), consistent with the female protective model. However, differences in recurrence rates between males from FC families and from families with only affected male children (MO) are not significant, nor are differences in recurrence rates between males and females within FC families, suggesting that simple differences in the magnitude of risk variant loads may not fully account for the differences in risk between MO and FC families. Instead, I propose two plausible modifications to the multiple threshold liability model: for one, it may be that high-risk families can be sub-classified to those in which the typical female-protective factors are compromised and those in which they are robust, or two, it may instead be that the specific variants carried by high-risk families can be sub-classified to those that are penetrant in females and those that are not. However, without full knowledge of each family's common and rare variant background, or information about potential sex-differential risk factors such as hormone levels, I cannot yet distinguish between these two possibilities. Furthermore, any extrapolation of the results observed here to the general population must be made with caution since AGRE families were recruited under specific ascertainment schemes that have likely introduced biases into the family

set available for analysis, such as a greater proportion of concordant twin pairs or less stoppage than is typical.

Second, I carried out a genetic linkage study on subsets of AGRE families defined by the sex of their affected children to identify genomic loci with sex-differential risk, with the eventual goal to use these loci as anchors for investigating the mechanistic details of sex-differential risk and protection. Published linkage analyses report numerous, different loci for overall and sex-differential ASD risk, likely due to small sample sizes and risk locus heterogeneity. Here, I test a large family sample derived from a single cohort, and I also stratify families into sex-defined subgroups for analysis, strategies that aim to reduce heterogeneity in the samples and enhance linkage signal. In the full family sample, I am able to replicate a previously identified linkage peak on chromosome 20p13. I also find several loci with suggestive sex-differential signal, and randomization testing suggests that LOD scores greater than or equal to those observed are unlikely to occur by chance at chromosomes 1p31.3 (MO), 8p21.2 (FC), and 8p12 (FC). Results from association testing within these regions are unable to identify the gene or genes within the regions that harbor risk variants, as I do not find any common, single-nucleotide variant alleles that are significantly associated with ASD status, though I am likely underpowered for even these regional analyses. Without identifying specific risk variants, I cannot conclude definitively if the relatively small linkage signals I identify are flagging sites of functional risk or if they are the result of statistical noise. I also cannot move forward with studies of risk and protective mechanisms at the molecular level around specific genes, as such genes remain to be identified. However, the fact that randomization testing suggests that the identified sex-differential loci are robust is still compelling, and further work to identify variants within these loci may be fruitful.

I have outlined strategies for identifying such variants, which include association testing for common SNPs and targeted sequencing to find rare variants that may cluster in particular genes across families. After targeted sequence capture and high throughput sequencing to identify variants, I have proposed a cost-effective, modified case-control paradigm for analysis, in which subjects from families with a high degree of allele sharing across the linkage region are considered as cases and subjects from families sharing no alleles are considered as controls. Power calculations show that these regional analyses are reasonably powered, though these calculations are based only on estimated observations of protein-disrupting single nucleotide variants. Other assumptions made for the power calculations include a low per-gene deleterious variant frequency that approaches the per-generation (*de novo*) rate of non-synonymous mutations, and the assumption of low locus heterogeneity within linkage regions, i.e. that deleterious variants from different families converge on one or just a few genes. Whether these assumptions are reasonable remains to be seen, as the specifics of the heritable component of ASD risk are largely unknown. Other goals of this study would therefore be exploratory in nature, to characterize patterns in risk variant class, location, and transmission in multiplex families. In the case that this approach identifies rare variants, shared between siblings and clustered in the same gene(s) across families, as intended, results would demonstrate a renewed utility for linkage analyses for identifying regions in which to look for familial risk variation in different sample sets.

Third, in addition to looking for novel risk loci that can serve as targets for work on the mechanistic details driving sex-differential ASD risk, I have also applied a different approach to investigate the relationship between known risk genes and sex differences in the typical brain. Specifically, I have assessed sex-differential gene expression patterns in human cortex from

adult and prenatal subjects, as sex differences in gene expression levels are likely to play a role in the organization or maintenance of a sexually dimorphic brain. By evaluating ASD risk genes, associated pathways, and cell type marker genes for evidence of sex-differential expression, I have found that known ASD risk genes are not expressed at significantly different levels in male and female cortex, though genes from a co-expression module up-regulated in ASD cortex in a previous study and astrocyte markers are expressed at higher levels in males. These gene expression patterns suggest that cortical astrocytes may be sexually dimorphic, and other work demonstrating functional interactions between glia and neuronal synapse formation and function further suggests that glia are well positioned to modulate the impact of variants in ASD risk genes at the cellular and circuit levels. Thus, sexually dimorphic astrocytes may lead to differences in cortical neuronal connectivity, and it is the operation of ASD risk genes within this sexually dimorphic circuitry that causes risk variants to have different effects in the male and female brain. These sex-differential effects may then translate to sex-differential presentation of the ASD phenotype.

I address several limitations to this work, the first of which is sample size, as these analyses were run on samples from a small number of individuals spanning wide age ranges, particularly for the adult data. Additionally, the sex differences in gene expression levels that I have observed are small in amplitude, suggesting that they may reflect subtle shifts in cortical anatomy or function. Without further work to identify the transcriptional regulatory mechanisms operating on the identified sex-differentially expressed genes, I also cannot determine from this analysis whether genes that are expressed at higher levels in males result from active up-regulation of these genes in males, from down-regulation in females, or a combination of both processes. Even so, these findings are intriguing in that they suggest that genes with sex-

differential expression interact with ASD risk genes at points downstream from the transcriptional regulation of ASD genes themselves. This model is consistent with the observation of stable sex-differential risk even in the face of extreme locus heterogeneity.

Findings from the work presented in this dissertation suggest several directions for future follow up. For one, as outlined in chapter 4, a thorough evaluation of genomic sequences within linkage regions to identify rare, deleterious variants will be critical to establish the functional significance of linkage peaks, and to identify loci for detailed study of sex-differential mechanisms and pathways. Characterization studies of risk variants identified specifically in MO or FC families along with investigation of factors related to sex-specific risk and protection in MO versus FC families will also allow for the evaluation of whether a family-based (families with or lacking female protective factors) or gene-based (variants penetrant or impervious in females) modification to the multiple threshold liability model is most correct. Also, as mentioned above, future studies of sex-differential gene expression in larger samples may be able to more definitively identify genes consistently and significantly differentially expressed by sex, even if the magnitude of the difference in expression level is small. Characterization of sex-differential gene expression in other brain regions, such as striatum, amygdala, or hypothalamic nuclei may also be informative and may reveal patterns not observed in neocortex. Along these lines, transcriptional profiling from single cells may also be useful for investigating the bases for the cell type differential expression patterns I have observed; for example, it may be that astrocytes show robust sex differences in gene expression levels, but that neurons do not.

Overall, across the studies described in this dissertation, I have applied several approaches to investigate the sex bias in ASD prevalence as it relates to genetic risk for ASD. Through this investigation, I have found that multiplex families from AGRE show evidence

consistent with predictions from the female protective model, suggesting that the prioritization of FC families for future genetic studies may enrich samples for more deleterious, detectable variants. I have applied this approach in a linkage analysis of families with and without an autistic female child and identify several suggestive, sex-differential loci, but no discernible common variant signal. For continued follow up of these loci, I propose a strategy for rare variant discovery and analysis that is directed by the location and family-level contributions to the linkage signals, with goals to both implicate specific genes for future targeted work on sex-specific modulatory mechanisms, and to characterize features of heritable risk variation for ASD. Results from my analysis of gene expression patterns in typical male and female cortex also suggest that ASD risk genes are not differentially regulated by sex, but that sex-specific risk and/or protective factors are more likely to act downstream from, or in pathways interacting with, risk variant-carrying genes. While much work remains to fully elucidate the mechanisms acting on the developing brain that modulate risk for the presentation of an ASD phenotype, the studies presented here bring to light several, potentially fundamental, features of sex-differential risk as related to genetics, and suggest promising directions for continued research in this area.

REFERENCES

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. Nature, 467(7319), 1061-1073. doi: 10.1038/nature09534

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet, 30*(1), 97-101. doi: 10.1038/ng786

Abrahams, B. S., & Geschwind, D. H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet, 9*(5), 341-355. doi: 10.1038/nrg2346

American Psychiatric Association., & American Psychiatric Association. DSM-5 Task Force. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5* (5th ed.). Washington, D.C.: American Psychiatric Association.

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. *bioRxiv*.

Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T. R., . . . Hallmayer, J. (2010). A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet, 19*(20), 4072-4082. doi: 10.1093/hmg/ddq307

Araque, A., Parpura, V., Sanzgiri, R. P., & Haydon, P. G. (1999). Tripartite synapses: glia, the unacknowledged partner. *Trends Neurosci, 22*(5), 208-215.

Asperger, H. (1944). The "autistic psychopathy" in childhood. *Archiv Fur Psychiatrie Und Nervenkrankheiten, 117*(1), 76-136. doi: 10.1007/Bf01837709

Attwood, T. (2006). Autism and Asperger syndrome: preparing for adulthood, 2nd edn. *Journal of Child Psychology and Psychiatry, 47*(2), 223A-224.

Auerbach, B. D., Osterweil, E. K., & Bear, M. F. (2011). Mutations causing syndromic autism define an axis of synaptic pathophysiology. *Nature, 480*(7375), 63-68. doi: 10.1038/nature10658

Auranen, M., Vanhala, R., Varilo, T., Ayers, K., Kempas, E., Ylisaukko-Oja, T., . . . Jarvela, I. (2002). A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. *Am J Hum Genet, 71*(4), 777-790. doi: 10.1086/342720

Auyeung, B., Baron-Cohen, S., Ashwin, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2009). Fetal testosterone and autistic traits. *Br J Psychol, 100*(Pt 1), 1-22. doi: 10.1348/000712608X311731

Auyeung, B., Baron-Cohen, S., Chapman, E., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Foetal testosterone and the child systemizing quotient. *European Journal of Endocrinology, 155*, S123-S130. doi: 10.1530/Eje.1.02260

Azcoitia, I., Sierra, A., & Garcia-Segura, L. M. (1999). Localization of estrogen receptor beta-immunoreactivity in astrocytes of the adult rat brain. *Glia, 26*(3), 260-267.

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. *Psychol Med, 25*(1), 63-77.

Bailey, D. B., Jr., Mesibov, G. B., Hatton, D. D., Clark, R. D., Roberts, J. E., & Mayhew, L. (1998). Autistic behavior in young boys with fragile X syndrome. *J Autism Dev Disord, 28*(6), 499-508.

Banach, R., Thompson, A., Szatmari, P., Goldberg, J., Tuff, L., Zwaigenbaum, L., & Mahoney, W. (2009). Brief Report: Relationship between non-verbal IQ and gender in autism. *J Autism Dev Disord, 39*(1), 188-193. doi: 10.1007/s10803-008-0612-4

Barbaresi, W., Katusic, S. K., Colligan, R. C., Pankratz, V. S., Weaver, A. L., Weber, K. J., . . . Jacobsen, S. J. (2002). How common is attention-deficit/hyperactivity disorder? Incidence in a population-based birth cohort in Rochester, Minn. *Archives of Pediatrics & Adolescent Medicine, 156*(3), 217-224.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends Cogn Sci, 6*(6), 248-254.

Baron-Cohen, S., Auyeung, B., Norgaard-Pedersen, B., Hougaard, D. M., Abdallah, M. W., Melgaard, L., . . . Lombardo, M. V. (2014). Elevated fetal steroidogenic activity in autism. *Mol Psychiatry*. doi: 10.1038/mp.2014.48

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philos Trans R Soc Lond B Biol Sci, 358*(1430), 361-374. doi: 10.1098/rstb.2002.1206

Baron-Cohen, S., & Wheelwright, S. (2003). The Friendship Questionnaire: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *J Autism Dev Disord, 33*(5), 509-517.

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J Autism Dev Disord, 34*(2), 163-175.

Barrett, S., Beck, J. C., Bernier, R., Bisson, E., Braun, T. A., Casavant, T. L., . . . et al. (1999). An autosomal genomic screen for autism. Collaborative linkage study of autism. *Am J Med Genet, 88*(6), 609-615.

Bartlett, C. W., Goedken, R., & Vieland, V. J. (2005). Effects of updating linkage evidence across subsets of data: reanalysis of the autism genetic resource exchange data set. *Am J Hum Genet, 76*(4), 688-695. doi: 10.1086/429345

Barzilai, N., Atzmon, G., Schechter, C., Schaefer, E. J., Cupples, A. L., Lipton, R., . . . Shuldiner, A. R. (2003). Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA, 290*(15), 2030-2040. doi: 10.1001/jama.290.15.2030

Basu, S. N., Kollu, R., & Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. *Nucleic Acids Res, 37*(Database issue), D832-836. doi: 10.1093/nar/gkn835

Berg, J. M., & Geschwind, D. H. (2012). Autism genetics: searching for specificity and convergence. *Genome Biol, 13*(7), 247. doi: 10.1186/gb4034

Betancur, C., Leboyer, M., & Gillberg, C. (2002). Increased rate of twins among affected sibling pairs with autism. *Am J Hum Genet, 70*(5), 1381-1383. doi: 10.1086/340364

Bezzi, P., Carmignoto, G., Pasti, L., Vesce, S., Rossi, D., Rizzini, B. L., . . . Volterra, A. (1998). Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. *Nature, 391*(6664), 281-285. doi: 10.1038/34651

Bishop, D. V., Jacobs, P. A., Lachlan, K., Wellesley, D., Barnicoat, A., Boyd, P. A., . . . Scerif, G. (2011). Autism, language and communication in children with sex chromosome trisomies. *Arch Dis Child, 96*(10), 954-959. doi: 10.1136/adc.2009.179747

Bolte, S., Duketis, E., Poustka, F., & Holtmann, M. (2011). Sex differences in cognitive domains and their clinical correlates in higher-functioning autism spectrum disorders. *Autism, 15*(4), 497-511. doi: 10.1177/1362361310391116

Bolton, P., Macdonald, H., Pickles, A., Rios, P., Goode, S., Crowson, M., . . . Rutter, M. (1994). A case-control family history study of autism. *J Child Psychol Psychiatry, 35*(5), 877-900.

Bouligand, J., Ghervan, C., Tello, J. A., Brailly-Tabard, S., Salenave, S., Chanson, P., . . . Young, J. (2009). Isolated familial hypogonadotropic hypogonadism and a GNRH1 mutation. *N Engl J Med, 360*(26), 2742-2748. doi: 10.1056/NEJMoa0900136

Bourgeron, T. (2009). A synaptic trek to autism. *Curr Opin Neurobiol, 19*(2), 231-234. doi: 10.1016/j.conb.2009.06.003

BrainSpan. (2013). BrainSpan: Atlas of the Developing Human Brain [Internet]. from http://www.brainspan.org

Brkanac, Z., Spencer, D., Shendure, J., Robertson, P. D., Matsushita, M., Vu, T., . . . Raskind, W. H. (2009). IFRD1 is a candidate gene for SMNA on chromosome 7q22-q23. *Am J Hum Genet, 84*(5), 692-697. doi: 10.1016/j.ajhg.2009.04.008

Brownlees, J., Ackerley, S., Grierson, A. J., Jacobsen, N. J., Shea, K., Anderton, B. H., . . . Miller, C. C. (2002). Charcot-Marie-Tooth disease neurofilament mutations disrupt neurofilament assembly and axonal transport. *Hum Mol Genet, 11*(23), 2837-2844.

Brugha, T. S., McManus, S., Bankart, J., Scott, F., Purdon, S., Smith, J., . . . Meltzer, H. (2011). Epidemiology of autism spectrum disorders in adults in the community in England. *Arch Gen Psychiatry, 68*(5), 459-465. doi: 10.1001/archgenpsychiatry.2011.38

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., . . . Barres, B. A. (2008). A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci, 28*(1), 264-278. doi: 10.1523/JNEUROSCI.4178-07.2008

Cantor, R. M., Kono, N., Duvall, J. A., Alvarez-Retuerto, A., Stone, J. L., Alarcon, M., . . . Geschwind, D. H. (2005). Replication of autism linkage: fine-mapping peak at 17q21. *Am J Hum Genet, 76*(6), 1050-1056. doi: 10.1086/430278

Carter, A. S., Black, D. O., Tewani, S., Connolly, C. E., Kadlec, M. B., & Tager-Flusberg, H. (2007). Sex differences in toddlers with autism spectrum disorders. *J Autism Dev Disord, 37*(1), 86-97. doi: 10.1007/s10803-006-0331-7

Chapman, E., Baron-Cohen, S., Auyeung, B., Knickmeyer, R., Taylor, K., & Hackett, G. (2006). Fetal testosterone and empathy: evidence from the empathy quotient (EQ) and the "reading the mind in the eyes" test. *Soc Neurosci, 1*(2), 135-148. doi: 10.1080/17470910600992239

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior & Development, 23*(1), 113-118. doi: 10.1016/S0163-6383(00)00032-1

Constantino, J. N. (2012). *Social Responsiveness Scale Manual*. Los Angeles: Western Psychological Services.

Constantino, J. N., & Todd, R. D. (2005). Intergenerational transmission of subthreshold autistic traits in the general population. *Biol Psychiatry, 57*(6), 655-660. doi: 10.1016/j.biopsych.2004.12.014

Constantino, J. N., Zhang, Y., Frazier, T., Abbacchi, A. M., & Law, P. (2010). Sibling recurrence and the genetic epidemiology of autism. *Am J Psychiatry, 167*(11), 1349-1356. doi: 10.1176/appi.ajp.2010.09101470

Coon, H., Matsunami, N., Stevens, J., Miller, J., Pingree, C., Camp, N. J., . . . McMahon, W. (2005). Evidence for linkage on chromosome 3q25-27 in a large autism extended pedigree. *Hum Hered, 60*(4), 220-226. doi: 10.1159/000090546

Creswell, C. S., & Skuse, D. H. (1999). Autism in association with Turner syndrome: Genetic implications for male vulnerability to pervasive developmental disorders. *Neurocase, 5*(6), 511-518. doi: 10.1080/13554799908402746

Dai, Y., Yang, Y., MacLeod, V., Yue, X., Rapraeger, A. C., Shriver, Z., . . . Sanderson, R. D. (2005). HSulf-1 and HSulf-2 are potent inhibitors of myeloma tumor growth in vivo. *J Biol Chem, 280*(48), 40066-40073. doi: 10.1074/jbc.M508136200

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y., Mele, A., Fraser, C. E., . . . Darnell, R. B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell, 146*(2), 247-261. doi: 10.1016/j.cell.2011.06.013

das Neves, L., Duchala, C. S., Tolentino-Silva, F., Haxhiu, M. A., Colmenares, C., Macklin, W. B., . . . Gronostajski, R. M. (1999). Disruption of the murine nuclear factor I-A gene (Nfia) results in perinatal lethality, hydrocephalus, and agenesis of the corpus callosum. *Proc Natl Acad Sci U S A, 96*(21), 11946-11951.

Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C. J., Gronostajski, R. M., & Anderson, D. J. (2006). The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. *Neuron, 52*(6), 953-968. doi: 10.1016/j.neuron.2006.11.019

Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators, C. f. D. C. a. P. (2012). Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR Surveill Summ, 61*(3), 1-19.

Developmental Disabilities Monitoring Network Surveillance Year 2010 Principal Investigators, C. f. D. C. a. P. (2014). Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2010. *MMWR Surveill Summ, 63*(2), 1-21.

Donnelly, S. L., Wolpert, C. M., Menold, M. M., Bass, M. P., Gilbert, J. R., Cuccaro, M. L., . . . Pericak-Vance, M. A. (2000). Female with autistic disorder and monosomy X (Turner syndrome): parent-of-origin effect of the X chromosome. *Am J Med Genet, 96*(3), 312-316.

Du, K., Herzig, S., Kulkarni, R. N., & Montminy, M. (2003). TRB3: a tribbles homolog that inhibits Akt/PKB activation by insulin in liver. *Science, 300*(5625), 1574-1577. doi: 10.1126/science.1079817

Dunn, L. M., Bulheller, S., & Häcker, H. (1965). *Peabody picture vocabulary test*. Circle Pines, MN: American Guidance Service.

Dworzynski, K., Ronald, A., Bolton, P., & Happe, F. (2012). How Different Are Girls and Boys Above and Below the Diagnostic Threshold for Autism Spectrum Disorders? *J Am Acad Child Adolesc Psychiatry, 51*(8), 788-797. doi: 10.1016/J.Jaac.2012.05.018

Elsmen, E., Pupp, I.H., & Hellstrom-Westas, L. (2004). Preterm male infants need more initial respiratory and circulatory support than female infants. *Acta Paediatrica, 93*(4), 529-533. doi: 10.1080/08035250410024998

Folstein, S., & Rutter, M. (1977). Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry, 18*(4), 297-321.

Folstein, S. E., & Piven, J. (1991). Etiology of autism: genetic influences. *Pediatrics, 87*(5 Pt 2), 767-773.

Fombonne, E. (1999). The epidemiology of autism: a review. *Psychol Med, 29*(4), 769-786.

Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatr Res, 65*(6), 591-598. doi: 10.1203/PDR.0b013e31819e7203

Freeman, R. D., Fast, D. K., Burd, L., Kerbeshian, J., Robertson, M. M., & Sandor, P. (2000). An international perspective on Tourette syndrome: selected findings from 3,500 individuals in 22 countries. *Dev Med Child Neurol, 42*(7), 436-447.

Garcia-Segura, L. M., Lorenz, B., & DonCarlos, L. L. (2008). The role of glia in the hypothalamus: implications for gonadal steroid feedback and reproductive neuroendocrine output. *Reproduction, 135*(4), 419-429. doi: 10.1530/REP-07-0540

Gecz, J., Shoubridge, C., & Corbett, M. (2009). The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet, 25*(7), 308-316. doi: 10.1016/j.tig.2009.05.002

Geier, D. A., & Geier, M. R. (2007). A prospective assessment of androgen levels in patients with autistic spectrum disorders: biochemical underpinnings and suggested therapies. *Neuro Endocrinol Lett, 28*(5), 565-573.

Geier, D. A., Kern, J. K., King, P. G., Sykes, L. K., & Geier, M. R. (2012). An evaluation of the role and treatment of elevated male hormones in autism spectrum disorders. *Acta Neurobiol Exp (Wars), 72*(1), 1-17.

Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L., & Spence, S.J. (2001). The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet, 69*(2), 463-466. doi: 10.1086/321292

Giarelli, E., Wiggins, L. D., Rice, C. E., Levy, S. E., Kirby, R. S., Pinto-Martin, J., & Mandell, D. (2010). Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. *Disability and Health Journal, 3*(2), 107-116. doi: 10.1016/J.Dhjo.2009.07.001

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

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., . . . Hakonarson, H. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature, 459*(7246), 569-573. doi: 10.1038/nature07953

Goin-Kochel, R. P., Abbacchi, A., Constantino, J. N., & Autism Genetic Resource Exchange, C. (2007). Lack of evidence for increased genetic loading for autism among families of affected females: a replication from family history data in two large samples. *Autism, 11*(3), 279-286. doi: 10.1177/1362361307076857

Greenberg, D. A., Hodge, S. E., Sowinski, J., & Nicoll, D. (2001). Excess of twins among affected sibling pairs with autism: implications for the etiology of autism. *Am J Hum Genet, 69*(5), 1062-1067. doi: 10.1086/324191

Gronborg, T. K., Schendel, D. E., & Parner, E. T. (2013). Recurrence of autism spectrum disorders in full- and half-siblings and trends over time: a population-based cohort study. *JAMA Pediatr, 167*(10), 947-953. doi: 10.1001/jamapediatrics.2013.2259

Gunnes, N., Suren, P., Bresnahan, M., Hornig, M., Lie, K. K., Lipkin, W. I., . . . Stoltenberg, C. (2013). Interpregnancy interval and risk of autistic disorder. *Epidemiology, 24*(6), 906-912. doi: 10.1097/01.ede.0000434435.52506.f5

Haber, M., Zhou, L., & Murai, K. K. (2006). Cooperative astrocyte and dendritic spine dynamics at hippocampal excitatory synapses. *J Neurosci, 26*(35), 8881-8891. doi: 10.1523/JNEUROSCI.1302-06.2006

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. *Arch Gen Psychiatry, 68*(11), 1095-1102. doi: 10.1001/archgenpsychiatry.2011.76

Hama, H., Hara, C., Yamaguchi, K., & Miyawaki, A. (2004). PKC signaling mediates global enhancement of excitatory synaptogenesis in neurons triggered by local contact with astrocytes. *Neuron, 41*(3), 405-415.

Hansen, K. D., Irizarry, R. A., & Wu, Z. J. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics, 13*(2), 204-216. doi: 10.1093/Biostatistics/Kxr054

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., . . . Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol, 7 Suppl 1*, S4 1-9. doi: 10.1186/gb-2006-7-s1-s4

Hattier, M. A., Matson, J. L., Tureck, K., & Horovitz, M. (2011). The effects of gender and age on repetitive and/or restricted behaviors and interests in adults with autism spectrum disorders and intellectual disability. *Res Dev Disabil, 32*(6), 2346-2351. doi: 10.1016/j.ridd.2011.07.028

Hoen, P. A., Friedlander, M. R., Almlof, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., . . . Lappalainen, T. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol, 31*(11), 1015-1022. doi: 10.1038/nbt.2702

Hoffmann, T. J., Windham, G. C., Anderson, M., Croen, L. A., Grether, J. K., & Risch, N. (2014). Evidence of Reproductive Stoppage in Families With Autism Spectrum Disorder: A Large, Population-Based Cohort Study. *JAMA Psychiatry*. doi: 10.1001/jamapsychiatry.2014.420

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi: 10.1371/journal.pgen.1000529

Huang, J., & Vieland, V. J. (2001). Comparison of 'model-free' and 'model-based' linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications. *Hum Hered, 51*(4), 217-225. doi: 53345

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., . . . Wigler, M. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron, 74*(2), 285-299. doi: 10.1016/j.neuron.2012.04.009

Jacquemont, S., Coe, B. P., Hersch, M., Duyzend, M. H., Krumm, N., Bergmann, S., . . . Eichler, E. E. (2014). A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am J Hum Genet, 94*(3), 415-425. doi: 10.1016/j.ajhg.2014.02.001

Jamain, S., Quach, H., Betancur, C., Rastam, M., Colineaux, C., Gillberg, I. C., . . . Paris Autism Research International Sibpair, S. (2003). Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nat Genet, 34*(1), 27-29. doi: 10.1038/ng1136

Jha, P., Sheth, D., & Ghaziuddin, M. (2007). Autism spectrum disorder and Klinefelter syndrome. *Eur Child Adolesc Psychiatry, 16*(5), 305-308. doi: 10.1007/s00787-007-0601-8

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127. doi: 10.1093/Biostatistics/Kxj037

Kang, H. J., Kawasawa, Y. I., Cheng, F., Zhu, Y., Xu, X., Li, M., . . . Sestan, N. (2011). Spatio-temporal transcriptome of the human brain. *Nature, 478*(7370), 483-489. doi: 10.1038/nature10523

Kanner, L. (1943). Autistic Disturbances of Affective Contact. *Nervous Child, 2*(3), 217-250.

Kennedy, R. B., Ovsyannikova, I. G., Pankratz, V. S., Haralambieva, I. H., Vierkant, R. A., & Poland, G. A. (2012). Genome-wide analysis of polymorphisms associated with cytokine responses in smallpox vaccine recipients. *Hum Genet, 131*(9), 1403-1421. doi: 10.1007/s00439-012-1174-2

Kieslinger, M., Folberth, S., Dobreva, G., Dorn, T., Croci, L., Erben, R., . . . Grosschedl, R. (2005). EBF2 regulates osteoblast-dependent differentiation of osteoclasts. *Dev Cell, 9*(6), 757-767. doi: 10.1016/j.devcel.2005.10.009

Kim, I. G., Lee, S. C., Lee, J. H., Yang, J. M., Chung, S. I., & Steinert, P. M. (1994). Structure and organization of the human transglutaminase 3 gene: evolutionary relationship to the transglutaminase family. *J Invest Dermatol, 103*(2), 137-142.

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. *Am J Psychiatry, 168*(9), 904-912. doi: 10.1176/appi.ajp.2011.10101532

Kiss-Toth, E., Bagstaff, S. M., Sung, H. Y., Jozsa, V., Dempsey, C., Caunt, J. C., . . . Dower, S. K. (2004). Human tribbles, a protein family controlling mitogen-activated protein kinase cascades. *J Biol Chem, 279*(41), 42703-42708. doi: 10.1074/jbc.M407732200

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. *Mol Autism, 3*(1), 9. doi: 10.1186/2040-2392-3-9

Knickmeyer, R., Baron-Cohen, S., Raggatt, P., & Taylor, K. (2005). Foetal testosterone, social relationships, and restricted interests in children. *J Child Psychol Psychiatry, 46*(2), 198-210. doi: 10.1111/j.1469-7610.2004.00349.x

Kohane, I. S., McMurry, A., Weber, G., MacFadden, D., Rappaport, L., Kunkel, L., . . . Churchill, S. (2012). The co-morbidity burden of children and young adults with autism spectrum disorders. *PLoS One, 7*(4), e33224. doi: 10.1371/journal.pone.0033224

Kong, A., & Cox, N. J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet, 61*(5), 1179-1188. doi: 10.1086/301592

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

Kong, X., Murphy, K., Raj, T., He, C., White, P. S., & Matise, T. C. (2004). A combined linkage-physical map of the human genome. *Am J Hum Genet, 75*(6), 1143-1148. doi: 10.1086/426405

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): an instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. *Res Dev Disabil, 32*(6), 2875-2888. doi: 10.1016/j.ridd.2011.05.017

Lai, D. C., Tseng, Y. C., Hou, Y. M., & Guo, H. R. (2012). Gender and geographic differences in the prevalence of autism spectrum disorders in children: analysis of data from the national disability registry of Taiwan. *Res Dev Disabil, 33*(3), 909-915. doi: 10.1016/j.ridd.2011.12.015

Lai, M. C., Lombardo, M. V., Pasco, G., Ruigrok, A. N., Wheelwright, S. J., Sadek, S. A., . . . Baron-Cohen, S. (2011). A behavioral comparison of male and female adults with high functioning autism spectrum conditions. *PLoS One, 6*(6), e20835. doi: 10.1371/journal.pone.0020835

Lajonchere, C.M., & the AGRE Consortium. (2010). Changing the landscape of autism research: the autism genetic resource exchange. *Neuron, 68*(2), 187-191. doi: 10.1016/j.neuron.2010.10.009

Lamb, J. A., Barnby, G., Bonora, E., Sykes, N., Bacchelli, E., Blasi, F., . . . Monaco, A. P. (2005). Analysis of IMGSAC autism susceptibility loci: evidence for sex limited and parent of origin specific effects. *J Med Genet, 42*(2), 132-137. doi: 10.1136/jmg.2004.025668

Lander, E., & Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet, 11*(3), 241-247. doi: 10.1038/ng1195-241

Levy, D., Ronemus, M., Yamrom, B., Lee, Y. H., Leotta, A., Kendall, J., . . . Wigler, M. (2011). Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron, 70*(5), 886-897. doi: 10.1016/j.neuron.2011.05.015

Lin, J. M., Kilman, V. L., Keegan, K., Paddock, B., Emery-Le, M., Rosbash, M., & Allada, R. (2002). A role for casein kinase 2alpha in the Drosophila circadian clock. *Nature, 420*(6917), 816-820. doi: 10.1038/nature01235

Liu, J., Nyholt, D. R., Magnussen, P., Parano, E., Pavone, P., Geschwind, D., . . . Gilliam, T. C. (2001). A genomewide screen for autism susceptibility loci. *Am J Hum Genet, 69*(2), 327-340.

Lombardo, M. V., Ashwin, E., Auyeung, B., Chakrabarti, B., Taylor, K., Hackett, G., . . . Baron-Cohen, S. (2012). Fetal testosterone influences sexually dimorphic gray matter in the human brain. *J Neurosci, 32*(2), 674-680. doi: 10.1523/JNEUROSCI.4389-11.2012

Lord, C., & Schopler, E. (1985). Differences in sex ratios in autism as a function of measured intelligence. *J Autism Dev Disord, 15*(2), 185-193.

Lu, A. T., & Cantor, R. M. (2012). Allowing for sex differences increases power in a GWAS of multiplex Autism families. *Mol Psychiatry, 17*(2), 215-222. doi: 10.1038/mp.2010.127

Mandy, W., Chilvers, R., Chowdhury, U., Salter, G., Seigal, A., & Skuse, D. (2012). Sex Differences in Autism Spectrum Disorder: Evidence from a Large Sample of Children and Adolescents. *J Autism Dev Disord, 42*(7), 1304-1313. doi: 10.1007/S10803-011-1356-0

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., . . . Scherer, S. W. (2008). Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet, 82*(2), 477-488. doi: 10.1016/j.ajhg.2007.12.009

Martin, L. A., & Horriat, N. L. (2012). The effects of birth order and birth interval on the phenotypic expression of autism spectrum disorder. *PLoS One, 7*(11), e51049. doi: 10.1371/journal.pone.0051049

Matise, T. C., Chen, F., Chen, W., De La Vega, F. M., Hansen, M., He, C., . . . Buyske, S. (2007). A second-generation combined linkage physical map of the human genome. *Genome Res, 17*(12), 1783-1786. doi: 10.1101/gr.7156307

Mattila, M. L., Kielinen, M., Linna, S. L., Jussila, K., Ebeling, H., Bloigu, R., . . . Moilanen, I. (2011). Autism spectrum disorders according to DSM-IV-TR and comparison with DSM-5 draft criteria: an epidemiological study. *J Am Acad Child Adolesc Psychiatry, 50*(6), 583-592 e511. doi: 10.1016/j.jaac.2011.04.001

Mayes, S. D., & Calhoun, S. L. (2011). Impact of IQ, age, SES, gender, and race on autistic symptoms. *Research in Autism Spectrum Disorders, 5*(2), 749-757. doi: 10.1016/J.Rasd.2010.09.002

McCarthy, M. M., Todd, B. J., & Amateau, S. K. (2003). Estradiol modulation of astrocytes and the establishment of sex differences in the brain. *Ann N Y Acad Sci, 1007*, 283-297.

McCauley, J. L., Li, C., Jiang, L., Olson, L. M., Crockett, G., Gainer, K., . . . Sutcliffe, J. S. (2005). Genome-wide and Ordered-Subset linkage analyses provide support for autism loci on 17q and 19p with evidence of phenotypic and interlocus genetic correlates. *BMC Med Genet, 6*, 1. doi: 10.1186/1471-2350-6-1

Mikhail, F. M., Lose, E. J., Robin, N. H., Descartes, M. D., Rutledge, K. D., Rutledge, S. L., . . . Carroll, A. J. (2011). Clinically relevant single gene or intragenic deletions encompassing critical neurodevelopmental genes in patients with developmental delay, mental retardation, and/or autism spectrum disorders. *Am J Med Genet A, 155A*(10), 2386-2396. doi: 10.1002/ajmg.a.34177

Miller, J. A., Horvath, S., & Geschwind, D. H. (2010). Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc Natl Acad Sci U S A, 107*(28), 12698-12703. doi: 10.1073/pnas.0914257107

Miller, J. A., Oldham, M. C., & Geschwind, D. H. (2008). A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J Neurosci, 28*(6), 1410-1420. doi: 10.1523/JNEUROSCI.4098-07.2008

Miller, J. A., Woltjer, R. L., Goodenbour, J. M., Horvath, S., & Geschwind, D. H. (2013). Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med, 5*(5), 48. doi: 10.1186/gm452

Mong, J. A., Kurzweil, R. L., Davis, A. M., Rocca, M. S., & McCarthy, M. M. (1996). Evidence for sexual differentiation of glia in rat brain. *Horm Behav, 30*(4), 553-562. doi: 10.1006/hbeh.1996.0058

Mong, J. A., & McCarthy, M. M. (1999). Steroid-induced developmental plasticity in hypothalamic astrocytes: implications for synaptic patterning. *J Neurobiol, 40*(4), 602-619.

Mong, J. A., & McCarthy, M. M. (2002). Ontogeny of sexually dimorphic astrocytes in the neonatal rat arcuate. *Brain Res Dev Brain Res, 139*(2), 151-158.

Morisset, C. E., Barnard, K. E., & Booth, C. L. (1995). Toddlers Language-Development - Sex-Differences within Social Risk. *Developmental Psychology, 31*(5), 851-865. doi: 10.1037//0012-1649.31.5.851

Muller, C. M. (1992). A role for glial cells in activity-dependent central nervous plasticity? Review and hypothesis. *Int Rev Neurobiol, 34*, 215-281.

Muller, C. M., & Best, J. (1989). Ocular dominance plasticity in adult cat visual cortex after transplantation of cultured astrocytes. *Nature, 342*(6248), 427-430. doi: 10.1038/342427a0

Nakanishi, K., Ida, M., Suzuki, H., Kitano, C., Yamamoto, A., Mori, N., . . . Taketani, S. (2006). Molecular characterization of a transport vesicle protein Neurensin-2, a homologue of Neurensin-1, expressed in neural cells. *Brain Res, 1081*(1), 1-8. doi: 10.1016/j.brainres.2006.01.085

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., . . . Daly, M. J. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature, 485*(7397), 242-245. doi: 10.1038/nature11011

Ng, S. B., Nickerson, D. A., Bamshad, M. J., & Shendure, J. (2010). Massively parallel sequencing and rare disease. *Hum Mol Genet, 19*(R2), R119-124. doi: 10.1093/hmg/ddq390

Nguyen, A., Rauch, T. A., Pfeifer, G. P., & Hu, V. W. (2010). Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. *FASEB J, 24*(8), 3036-3051. doi: 10.1096/fj.10-154484

Niemi, M., Ikonen, M., & Hervonen, A. (1967). Histochemistry and fine structure of the interstitial tissues in the human fetal testis. In G. Wolstenholme (Ed.), *Ciba Foundation Colloquia on Endocrinology, Vol 16: The Testis* (pp. 31-55). London: Churchill.

Nikopoulos, K., Gilissen, C., Hoischen, A., van Nouhuys, C. E., Boonstra, F. N., Blokland, E. A., . . . Collin, R. W. (2010). Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet, 86*(2), 240-247. doi: 10.1016/j.ajhg.2009.12.016

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., . . . Eichler, E. E. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature, 485*(7397), 246-250. doi: 10.1038/nature10989

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., & Geschwind, D. H. (2008). Functional organization of the transcriptome in human brain. *Nat Neurosci, 11*(11), 1271-1282. doi: 10.1038/nn.2207

Oldham, M. C., Langfelder, P., & Horvath, S. (2012). Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst Biol, 6*, 63. doi: 10.1186/1752-0509-6-63

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., . . . Stone, W. L. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics, 128*(3), e488-495. doi: 10.1542/peds.2010-2825

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), 1008-1021. doi: 10.1016/j.cell.2013.10.031

Perea, G., Navarrete, M., & Araque, A. (2009). Tripartite synapses: astrocytes process and control synaptic information. *Trends Neurosci, 32*(8), 421-431. doi: 10.1016/j.tins.2009.05.001

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., . . . Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature, 466*(7304), 368-372. doi: 10.1038/nature09146

Piven, J., Palmer, P., Jacobi, D., Childress, D., & Arndt, S. (1997). Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. *Am J Psychiatry, 154*(2), 185-190.

Pohl, A., Cassidy, S., Auyeung, B., & Baron-Cohen, S. (2014). Uncovering steroidopathy in women with autism: a latent class analysis. *Mol Autism, 5*, 27. doi: 10.1186/2040-2392-5-27

Polanczyk, G., de Lima, M. S., Horta, B. L., Biederman, J., & Rohde, L. A. (2007). The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am J Psychiatry, 164*(6), 942-948. doi: 10.1176/appi.ajp.164.6.942

Price, W. H., & van der Molen, H. J. (1970). Plasma testosterone levels in males with the 47,XYY karyotype. *J Endocrinol, 47*(1), 117-122.

Purcell, S., Cherny, S. S., & Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics, 19*(1), 149-150.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Rabinowicz, T., Dean, D. E., Petetot, J. M., & de Courten-Myers, G. M. (1999). Gender differences in the human cerebral cortex: more neurons in males; more processes in females. *J Child Neurol, 14*(2), 98-107.

Ratcliffe, S. G., Read, G., Pan, H., Fear, C., Lindenbaum, R., & Crossley, J. (1994). Prenatal testosterone levels in XXY and XYY males. *Horm Res, 42*(3), 106-109.

Raven, J. C., & De Lemos, M. M. (1958). *Standard progressive matrices*. London: Lewis.

Rehman, A. U., Morell, R. J., Belyantseva, I. A., Khan, S. Y., Boger, E. T., Shahzad, M., . . . Friedman, T. B. (2010). Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet, 86*(3), 378-388. doi: 10.1016/j.ajhg.2010.01.030

Reich, R., Cloninger, C. R., & Guze, S. B. (1975). The multifactorial model of disease transmission: I. Description of the model and its use in psychiatry. *Br J Psychiatry, 127*, 1-10.

Ripke, S., Sanders, A. R., Kendler, K. S., Levinson, D. F., Sklar, P., Holmans, P. A., . . . Gejman, P. V. (2011). Genome-wide association study identifies five new schizophrenia loci. *Nat Genet, 43*(10), 969-976. doi: 10.1038/ng.940

Ripke, S., Wray, N. R., Lewis, C. M., Hamilton, S. P., Weissman, M. M., Breen, G., . . . Sullivan, P. F. (2013). A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry, 18*(4), 497-511. doi: 10.1038/mp.2012.21

Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F., & Ronald, A. (2013). Examining and interpreting the female protective effect against autistic behavior. *Proc Natl Acad Sci U S A, 110*(13), 5258-5262. doi: 10.1073/pnas.1211070110

Ronemus, M., Iossifov, I., Levy, D., & Wigler, M. (2014). The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet, 15*(2), 133-141. doi: 10.1038/nrg3585

Ross, J. L., Roeltgen, D. P., Kushner, H., Zinn, A. R., Reiss, A., Bardsley, M. Z., . . . Tartaglia, N. (2012). Behavioral and social phenotypes in boys with 47,XYY syndrome or 47,XXY Klinefelter syndrome. *Pediatrics, 129*(4), 769-778. doi: 10.1542/peds.2011-0719

Ruta, L., Ingudomnukul, E., Taylor, K., Chakrabarti, B., & Baron-Cohen, S. (2011). Increased serum androstenedione in adults with autism spectrum conditions. *Psychoneuroendocrinology, 36*(8), 1154-1163. doi: 10.1016/j.psyneuen.2011.02.007

Rutter, M. (1968). Concepts of autism: a review of research. *J Child Psychol Psychiatry, 9*(1), 1-25.

Sakai, Y., Shaw, C. A., Dawson, B. C., Dugas, D. V., Al-Mohtaseb, Z., Hill, D. E., & Zoghbi, H. Y. (2011). Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med, 3*(86), 86ra49. doi: 10.1126/scitranslmed.3002166

Sakuma, S., Tokuhara, D., Hattori, H., Matsuoka, O., & Yamano, T. (2009). Expression of estrogen receptor alpha and beta in reactive astrocytes at the male rat hippocampus after status epilepticus. *Neuropathology, 29*(1), 55-62. doi: 10.1111/j.1440-1789.2008.00946.x

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., . . . State, M. W. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron, 70*(5), 863-885. doi: 10.1016/j.neuron.2011.05.002

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., . . . State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature, 485*(7397), 237-241. doi: 10.1038/nature10945

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C. M., & Reichenberg, A. (2014). The familial risk of autism. *JAMA, 311*(17), 1770-1777. doi: 10.1001/jama.2014.4144

Sarachana, T., Xu, M., Wu, R. C., & Hu, V. W. (2011). Sex hormones in autism: androgens and estrogens differentially and reciprocally regulate RORA, a novel candidate gene for autism. *PLoS One, 6*(2), e17116. doi: 10.1371/journal.pone.0017116

Sato, D., Lionel, A. C., Leblond, C. S., Prasad, A., Pinto, D., Walker, S., . . . Scherer, S. W. (2012). SHANK1 Deletions in Males with Autism Spectrum Disorder. *Am J Hum Genet, 90*(5), 879-887. doi: 10.1016/j.ajhg.2012.03.017

Schellenberg, G. D., Dawson, G., Sung, Y. J., Estes, A., Munson, J., Rosenthal, E., . . . Wijsman, E. M. (2006). Evidence for multiple loci from a genome scan of autism kindreds. *Mol Psychiatry, 11*(11), 1049-1060, 1979. doi: 10.1038/sj.mp.4001874

Scheuerle, A., & Wilson, K. (2011). PARK2 copy number aberrations in two children presenting with autism spectrum disorder: further support of an association and possible evidence for a new microdeletion/microduplication syndrome. *Am J Med Genet B Neuropsychiatr Genet, 156B*(4), 413-420. doi: 10.1002/ajmg.b.31176

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., . . . Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol, 7*, 3. doi: 10.1186/1471-2199-7-3

Schwarz, E., Guest, P. C., Rahmoune, H., Wang, L., Levin, Y., Ingudomnukul, E., . . . Bahn, S. (2011). Sex-specific serum biomarker patterns in adults with Asperger's syndrome. *Mol Psychiatry, 16*(12), 1213-1220. doi: 10.1038/mp.2010.102

Schwarz, J. M., & Bilbo, S. D. (2012). Sex, glia, and development: interactions in health and disease. *Horm Behav, 62*(3), 243-253. doi: 10.1016/j.yhbeh.2012.02.018

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., . . . Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science, 316*(5823), 445-449. doi: 10.1126/science.1138659

Sham, P. C., Lin, M. W., Zhao, J. H., & Curtis, D. (2000). Power comparison of parametric and nonparametric linkage tests in small pedigrees. *Am J Hum Genet, 66*(5), 1661-1668. doi: 10.1086/302888

Shao, Y., Raiford, K. L., Wolpert, C. M., Cope, H. A., Ravan, S. A., Ashley-Koch, A. A., . . . Pericak-Vance, M. A. (2002). Phenotypic homogeneity provides increased support for

linkage on chromosome 2 in autistic disorder. *Am J Hum Genet, 70*(4), 1058-1061. doi: 10.1086/339765

Sklar, P., Ripke, S., Scott, L., Andreassen, O., Cichon, S., Craddock, N., . . . Purcell, S. (2011). Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet, 43*(10), 977-983. doi: 10.1038/ng.943

Skuse, D. H. (2000). Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. *Pediatr Res, 47*(1), 9-16.

Skuse, D. H., James, R. S., Bishop, D. V., Coppin, B., Dalton, P., Aamodt-Leeper, G., . . . Jacobs, P. A. (1997). Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. *Nature, 387*(6634), 705-708. doi: 10.1038/42706

Smail, P. J., Reyes, F. I., Winter, J. S. D., & Faiman, C. (1981). The Fetal Hormonal Environment and its Effect on the Morphogenesis of the Genital System. In S. J. Kogan & E. S. E. Hafez (Eds.), *Pediatric Andrology* (Vol. 7, pp. 9-19): Springer Netherlands.

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 397-420): Springer New York.

Solomon, M., Miller, M., Taylor, S. L., Hinshaw, S. P., & Carter, C. S. (2012). Autism Symptoms and Internalizing Psychopathology in Girls and Boys with Autism Spectrum Disorders. *J Autism Dev Disord, 42*(1), 48-59. doi: 10.1007/S10803-011-1215-Z

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales: Interview edition, survey form manual.* Circle Pines, MN: American Guidance Service.

Steen, E.E., Kallen, K., Marsal, K., Norman, M., & Hellstrom-Westas, L. (2014). Impact of sex on perinatal mortality and morbidity in twins. *Journal of Perinatal Medicine, 42*(2), 225-231. doi: 10.1515/Jpm-2013-0147

Stefansson, H., Sarginson, J., Kong, A., Yates, P., Steinthorsdottir, V., Gudfinnsson, E., . . . St Clair, D. (2003). Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. *Am J Hum Genet, 72*(1), 83-87.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., . . . Stefansson, K. (2002). Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet, 71*(4), 877-892. doi: 10.1086/342734

Stevenson, D.K., Verter, J., Fanaroff, A.A., Oh, W., Ehrenkranz, R.A., Shankaran, S., . . . Papile, L-A. (2000). Sex differences in outcomes of very low birthweight infants: the newborn male disadvantage. *Archives of Disease in Childhood, 83*(3), F182-F185. doi: 10.1136/Fn.83.3.F182

Stone, J. L., Merriman, B., Cantor, R. M., Yonan, A. L., Gilliam, T. C., Geschwind, D. H., & Nelson, S. F. (2004). Evidence for sex-specific risk alleles in autism spectrum disorder. *Am J Hum Genet, 75*(6), 1117-1123. doi: 10.1086/426034

Szatmari, P., Liu, X. Q., Goldberg, J., Zwaigenbaum, L., Paterson, A. D., Woodbury-Smith, M., . . . Thompson, A. (2012). Sex differences in repetitive stereotyped behaviors in autism: Implications for genetic liability. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics, 159B*(1), 5-12. doi: 10.1002/Ajmg.B.31238

Szatmari, P., Offord, D. R., & Boyle, M. H. (1989). Ontario Child Health Study - Prevalence of Attention Deficit Disorder with Hyperactivity. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 30*(2), 219-230. doi: 10.1111/J.1469-7610.1989.Tb00236.X

Szatmari, P., Paterson, A. D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X. Q., . . . Meyer, K. J. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet, 39*(3), 319-328. doi: 10.1038/ng1985

Takagishi, H., Takahashi, T., Yamagishi, T., Shinada, M., Inukai, K., Tanida, S., . . . Kameda, T. (2010). Salivary testosterone levels and autism-spectrum quotient in adults. *Neuro Endocrinol Lett, 31*(6), 837-841.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *J Speech Lang Hear Res, 40*(6), 1245-1260.

Trabzuni, D., Ramasamy, A., Imran, S., Walker, R., Smith, C., Weale, M. E., . . . North American Brain Expression, C. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. *Nat Commun, 4*, 2771. doi: 10.1038/ncomms3771

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics, 25*(9), 1105-1111. doi: 10.1093/bioinformatics/btp120

Trikalinos, T. A., Karvouni, A., Zintzaras, E., Ylisaukko-oja, T., Peltonen, L., Jarvela, I., & Ioannidis, J. P. (2006). A heterogeneity-based genome search meta-analysis for autism-spectrum disorders. *Mol Psychiatry, 11*(1), 29-36. doi: 10.1038/sj.mp.4001750

Tropeano, M., Ahn, J. W., Dobson, R. J., Breen, G., Rucker, J., Dixit, A., . . . Collier, D. A. (2013). Male-biased autosomal effect of 16p13.11 copy number variation in neurodevelopmental disorders. *PLoS One, 8*(4), e61365. doi: 10.1371/journal.pone.0061365

Tsai, L., Stewart, M. A., & August, G. (1981). Implication of sex differences in the familial transmission of infantile autism. *J Autism Dev Disord, 11*(2), 165-173.

van Rijn, S., Bierman, M., Bruining, H., & Swaab, H. (2012). Vulnerability for autism traits in boys and men with an extra X chromosome (47,XXY): the mediating role of cognitive flexibility. *J Psychiatr Res, 46*(10), 1300-1306. doi: 10.1016/j.jpsychires.2012.06.004

Viding, E., Spinath, F. M., Price, T. S., Bishop, D. V. M., Dale, P. S., & Plomin, R. (2004). Genetic and environmental influence on language impairment in 4-year-old same-sex and opposite-sex twins. *Journal of Child Psychology and Psychiatry, 45*(2), 315-325. doi: 10.1111/J.1469-7610.2004.00223.X

Vieland, V. J. (2006). Thermometers: something for statistical geneticists to think about. *Hum Hered, 61*(3), 144-156. doi: 10.1159/000093775

Vieland, V. J. (2011). Where's the evidence? *Hum Hered, 71*(1), 59-66. doi: 10.1159/000324838

Vieland, V. J., Wang, K., & Huang, J. (2001). Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data. *Hum Hered, 51*(4), 199-208. doi: 53343

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), 380-384. doi: 10.1038/nature10110

Volkmar, F. R., Szatmari, P., & Sparrow, S. S. (1993). Sex differences in pervasive developmental disorders. *J Autism Dev Disord, 23*(4), 579-591.

Wang, K., Zhang, H., Ma, D., Bucan, M., Glessner, J. T., Abrahams, B. S., . . . Hakonarson, H. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature, 459*(7246), 528-533. doi: 10.1038/nature07999

Weiss, L. A., Arking, D. E., Daly, M. J., & Chakravarti, A. (2009). A genome-wide linkage and association scan reveals novel loci for autism. *Nature, 461*(7265), 802-808. doi: 10.1038/nature08490

Werling, D. M., & Geschwind, D. H. (2013). Sex differences in autism spectrum disorders. *Curr Opin Neurol, 26*(2), 146-153. doi: 10.1097/WCO.0b013e32835ee548

Werling, D. M., Lowe, J. K., Luo, R., Cantor, R. M., & Geschwind, D. H. (2014). Replication of linkage at chromosome 20p13 and identification of suggestive sex-differential risk loci for autism spectrum disorder. *Mol Autism, 5*(1), 13. doi: 10.1186/2040-2392-5-13

Whitacre, C. C. (2001). Sex differences in autoimmune disease. *Nature Immunology, 2*(9), 777-780. doi: 10.1038/Ni0901-777

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

Winden, K. D., Oldham, M. C., Mirnics, K., Ebert, P. J., Swan, C. H., Levitt, P., . . . Geschwind, D. H. (2009). The organization of the transcriptional network in specific neuronal classes. *Mol Syst Biol, 5*, 291. doi: 10.1038/msb.2009.46

Yang, X., Schadt, E. E., Wang, S., Wang, H., Arnold, A. P., Ingram-Drake, L., . . . Lusis, A. J. (2006). Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res, 16*(8), 995-1004. doi: 10.1101/gr.5217506

Yeargin-Allsopp, M., Rice, C., Karapurkar, T., Doernberg, N., Boyle, C., & Murphy, C. (2003). Prevalence of autism in a US metropolitan area. *JAMA, 289*(1), 49-55.

Yokoyama, S., Woods, S. L., Boyle, G. M., Aoude, L. G., MacGregor, S., Zismann, V., . . . Brown, K. M. (2011). A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature, 480*(7375), 99-103. doi: 10.1038/nature10630

Yonan, A. L., Alarcon, M., Cheng, R., Magnusson, P. K., Spence, S. J., Palmer, A. A., . . . Gilliam, T. C. (2003). A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet, 73*(4), 886-897.

Yonan, A. L., Palmer, A. A., Smith, K. C., Feldman, I., Lee, H. K., Yonan, J. M., . . . Gilliam, T. C. (2003). Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. *Genes Brain Behav, 2*(5), 303-320.

Zhang, K., Lindsberg, P. J., Tatlisumak, T., Kaste, M., Olsen, H. S., & Andersson, L. C. (2000). Stanniocalcin: A molecular guard of neurons during cerebral ischemia. *Proc Natl Acad Sci U S A, 97*(7), 3637-3642. doi: 10.1073/pnas.070045897

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. *Proc Natl Acad Sci U S A, 104*(31), 12831-12836. doi: 10.1073/pnas.0705803104

Zwaigenbaum, L., Bryson, S. E., Szatmari, P., Brian, J., Smith, I. M., Roberts, W., . . . Roncadin, C. (2012). Sex differences in children with autism spectrum disorder identified within a high-risk infant cohort. *J Autism Dev Disord, 42*(12), 2585-2596. doi: 10.1007/s10803-012-1515-y