IDENTIFYING BIOLOGICAL PATHWAYS IMPLICATED IN DEFINED SUBGROUPS OF PHENOTYPIC EXPRESSION FOR AUTISM SPECTRUM DISORDERS

&

EVALUATING SMALL MOLECULE EFFECTS ON EXPRESSION OF *ASMT*

By

Olivia Jean Veatch

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

HUMAN GENETICS

December, 2013

Nashville, TN

Approved:	Date:
Jeremy Veenstra-VanderWeele	08/09/13
Douglas P. Mortlock	08/09/13
Colleen Niswender	08/09/13
Jonathan L. Haines	08/09/13
Tricia A. Thornton-Wells	08/09/13

To all of those affected by developmental disorders:

Patients, Clinicians, Teachers, Families & Friends

ACKNOWLEDGEMENTS

The work presented in this dissertation was supported by CTSA award UL1TR000445 to Olivia J. Veatch from the National Center for Advancing Translational Sciences and the Vanderbilt Program in Human Genetics.

First and foremost, I would like to express my sincere gratitude to Dr. Jonathan L. Haines for giving me the opportunity to accomplish this body of research and for being such a supportive mentor. I will always appreciate your faith in me and my ideas. Your example has inspired me to strive to be a better scientist every day. I would also like to thank my dissertation committee: Dr. Tricia A. Thornton-Wells (chair) for her expertise in multivariate statistics and complex human disease genetic analysis, Dr. Douglas P. Mortlock for his help with designing and implementing the molecular experiments, Dr. Colleen M. Niswender for her extensive knowledge of molecular drug development, and Dr. Jeremy M. Veenstra-VanderWeele for his vast knowledge of Autism Spectrum Disorders phenotypes, as well as, the molecular biology underlying the disorders. This work was significantly enhanced by your expert advice.

I am thankful to all the members of the Haines lab, past and present. I'd especially like to thank Dr. Nathalie Schnetz-Boutaud, Ping Mayo, M.S., and Melissa Allen for sharing with me their broad range of experience. I am especially grateful to Ping Mayo, who generated all cell lines used in these studies. Special thanks to Dr. Brian Yaspan for his invaluable help on the pathway-based genetic analyses. Thanks to my fellow mentorees, Dr. Jessica Cooke-Bailey, Dr. Anna C. Cummings, Laura D'Aoust, Mary F. Davis, M.S., Joshua Hoffman, M.S., and Dr. Rebecca Zuvich for offering support and advice whenever it was needed.

I would like to thank the Interdisciplinary Graduate Program (IGP) and the program in Human Genetics at Vanderbilt University. I'm especially grateful to my Directors of

Graduate Studies, Dr. Scott M. Williams and Dr. David Samuels, and Program Coordinator, Maria Ritchie for all they have done for me over the years. Thank you to the investigators and administrators in the Center for Human Genetics Research for all of their learning, and advice, during my graduate training. I would specifically like to thank Dr. Deborah Murdock for offering her time and knowledge whenever I needed it most. Thank you to Angelika Parl for maintaining the cell culture hood. Thank you to Alicia Davis and Kimberly Turner for helping with the logistics for numerous grant applications. Thank you to Julia Bayles for help formatting and submitting manuscripts. Thank you to the Computational Genomics Core, particularly Melissa Potter and Wayne Perkins, for helping guide and troubleshoot the computational work presented. Thank you to the DNA Core, especially Paxton Baker, M.S. for helping troubleshoot quantitative real-time PCR, and Lindsay Williamson, for access and storage of cell lines. I would also like to thank my fellow Human Genetics and IGP students for all of their support along the way.

Special thanks to the families who participated in the studies from which this work was made possible. We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange Consortium (AGRE) and Autism Genome Project (AGP) Consortium. AGRE is a program of Autism Speaks and is supported, in part, by grant 1U24MH081810 from the National Institute of Mental Health to C.M. Lajonchere (PI), and formerly by grant MH64547 to D.H. Geschwind (PI). AGP is funded by Autism Speaks, the Medical Research Council (UK), and the Health Research Board (Ireland).

I would like to personally express gratitude to my family and friends who have wholly supported me throughout my life and in my career thus far. Special thanks to my amazing father and mother, Richard and Shirley Veatch, my incredibly supportive husband, James M. Coyle, Jr, my loving sister and closest friend, Audra Veatch Bryson, my great friend and personal advisor, Tracy Macko Harris, and Dr. Zohreh Talebizadeh for taking a chance all those years ago on a very naïve undergrad.

TABLE OF CONTENTS

Page

Chapter

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

History of Autism Spectrum Disorders

Autism spectrum disorders (ASD, OMIM 209850) are complex, neurodevelopmental disorders characterized by impairments in social communication and the presence of restricted and repetitive behavioral patterns[13]. Autism was first described in 1943 by Dr. Leo Kanner[102]. Dr. Kanner described 11 patients, mostly boys, having a combination of severe social dysfunction, variable communication deficits, and the presence of repetitive restrictive behaviors. Interesting observations based on these initial case studies included the identification of large head size in approximately half of the subjects. Dr. Kanner also postulated a biological, genetic basis for the disorder. However, it was not until much later that autism began to be considered a distinct disorder in psychiatric diagnostic manuals. Since then, prevalence estimates have steadily been increasing with current estimates in the United States as high as 1 in 88 children[1]. These estimates vary widely across all sites, by sex (ASD are estimated to be almost 5 times more common among boys), and by racial/ethnic group.

There are numerous possible explanations for the substantial increase in ASD prevalence over such a short period of time. One is that the concept of autism has broadened from what was previously considered a 'strict' diagnosis of autistic disorder, to include individuals of normal intelligence with adequate language skills (DSM-IV Asperger Disorder), those not quite meeting diagnostic criteria in all three domains (DSM-IV Pervasive Developmental Disorders-Not Otherwise Specified), and those who develop normally for a period of time followed by regression in skills or a series of regressions in skills (DSM-IV Childhood Disintegrative Disorder)[12]. It is notable these diagnoses are not based on etiology, but on expert observation and assessment of behavioral and cognitive characteristics. How these clinical domains relate to underlying dysfunction in specific cognitive domains is essentially unknown.

Even within more unified diagnostic definitions, the severity of clinical presentation is quite heterogeneous. Some affected individuals also present with various comorbidities (i.e. epilepsy, mental retardation), endophenotypes (i.e. presence of savant skills, specific language impairment), and biomarkers (i.e. macrocephaly, hyperserotonemia)[62, 76, 151, 178, 209]. Thus, autistic disorder appears to be not a single entity but rather a complex phenotype expressing a continuum of symptom severity and neurocognitive impairments. This is reflected in the recent change in diagnostic criteria for ASD between the DSM-IV and the DSM-5. These revisions were motivated by the lack of empirical data supporting separate disorders within the autism spectrum[13].

Genetics of Autism Spectrum Disorders

ASD was also for a long time not considered to have any underlying genetic basis[81]. The first evidence for an inherited genetic component to autism came from twin studies published in 1977[60, 61]. These initial twin studies demonstrated a genetic susceptibility to the disorder and provided substantial evidence supporting biological origins. To date, there is overwhelming evidence suggesting strong genetic susceptibility factors underlying ASD. The sibling recurrence risk is estimated at 45–90 times greater than the population risk. Current estimates from twin studies indicate 58-60% of monozygotic twins are concordant for the full syndrome and 50-90% are concordant for related social or cognitive abnormalities[21, 45, 79]. There are also a number of syndromes with well-defined genetic causes associated with ASD. These include, but are not limited to, Rett syndrome, tuberous sclerosis, neurofibromatosis, and Fragile X Syndrome[14, 48, 104, 228]. The hallmark presentations of these syndromes are more homogeneous profiles of characteristic physical features, neurological impairment, and ASD symptoms. However, only a very small percentage of individuals with ASD (<1%) have an identifiable genetic etiology known to cause these monogenic disorders[3].

The reported prevalence and heritable nature of ASD suggests that genetic variation present at relatively common frequencies in the overall population contribute to the genetic etiology underlying these disorders. Numerous studies have evaluated the involvement of common variation in ASD. Results from these studies implicate a number of commonly occurring variations, across the genome, each with relatively small effect sizes[15, 17, 106, 133, 221]. It is hypothesized that many idiopathic ASD cases, those with no diagnosed clinical syndrome, are a result of the interactions of multiple common variants, each with small to moderate effect sizes. Identifying common variation with any appreciable influence on ASD risk has proven difficult; however, this is not incredibly surprising, given the obvious complexity of ASD. Common variation associated and/or linked to ASD is discussed in greater detail in Chapter III.

A large number of rare, recurrent, and non-recurrent mutations have been identified that are thought to lead to ASD[34, 142, 162]. Most of the identified rare mutations are small regions of chromosomal structural variation known as copy number variants (CNVs). Many of these CNVs have large effect sizes and some appear to be sufficient to cause ASD. Identified inherited CNVs, like those at 16p11 and 15q11-13, are transmitted from apparently unaffected parents, who may display some level of autistic traits, to affected offspring[176]. However, most identified CNVs are *de novo* events, arising in the germline. These *de novo* CNVs are reported in ~5–10% of ASD probands[25, 156, 176, 177]. Overall, CNVs are linked to a broad variety of clinical features, including severe neurological symptoms, severe ASD, milder autism-spectrum traits, and behavioral disorders outside of the autism spectrum[160]. Many CNVs found in ASD patients have also been found in patients specifically with intellectual disability and schizophrenia, but no ASD[47, 162].

Phenotypic heterogeneity characterizing CNV expressivity makes it difficult to determine whether an identified CNV is the sole cause of autism, confers vulnerability to the disease, or represents a chance finding. It is also important to note that many *de novo* CNVs associated with ASD, while rare, are also observed in unaffected controls, suggesting these variations are not necessarily causal or fully penetrant[34]. Some CNVs may be acting as complex genetic risk factors, with intermediate effect sizes, variable penetrance and variable expressivity[70].

The current results from numerous genetic analyses in ASD all indicate an incredible complexity of underlying genetic mechanisms. However, the known biological functions for recurrently implicated genes suggest involvement of shared molecular pathways. For example, numerous genes have been identified that encode proteins important to synaptic function. These include neurologins and neurexins, specifically *NLGN3, NLGN4* and *NRXN1*[58, 98]. Interactions between neuroligins and neurexins trigger the formation of functional pre-synaptic boutons[46]. Also included are post-synaptic scaffolding proteins, specifically *SHANK1*, *SHANK2*, and *SHANK3*[27, 203].

Another convergent molecular mechanism in ASD is related to morphogenesis. Numerous protein-altering mutations and cytogenetic abnormalities have been identified that affect morphogenetic and growth-regulating genes. These genes include *HOXA1*, the first HOX gene to be expressed during embryogenesis which is necessary for the proper development of the brainstem, cerebellum and several cranial nerves[42, 136, 208]. Another implicated growth-regulating gene is *EIF4E*, the rate limiting component of eukaryotic translation initiation that plays a key role in learning and memory[154]. Finally, mutations disrupting the tumor suppressor gene*, PTEN*, have been identified in numerous patients with ASD. Most subjects with autism carrying PTEN mutations are characterized by severe to extreme macrocephaly[35].

A collection of recent genetic evidence suggests that some ASD cases may result from abnormal Ca2+ homeostasis during neurodevelopment[112]. Several genetic studies have identified autism-related genes encoding ion channels, receptors, and Ca2+-regulated signaling proteins, often times crucial to central nervous system development. These genes include, *CACNA1C*, *CACNA1F*, *CACNA1H*, *KCNMA1*, *and SCN2A*[85, 117, 198, 199, 222].

Finally, the most consistently replicated genes harboring common variants related to ASD are: the *SLC6A4* gene encoding the serotonin transporter, the *EN2* gene, encoding the engrailed homeobox 2 protein (implicated in pattern formation during central nervous system development), the *OXTR* gene, encoding a G-protein coupled oxytocin receptor, the *CNTNAP2* gene, encoding a neurexin family protein that functions in the nervous system, the *GABRB3* gene, encoding a ligand-gated gamma-aminobutyric acid receptor, the *RELN* gene, encoding an extracellular matrix protein important for neuronal migration during development, the *ITGB3* gene, encoding an integrin important in cell adhesion and signaling, and the *MET* gene, encoding the Met proto-oncogene involved in brain development[160].

Small Molecule Compound Treatment of Autism Spectrum Disorders

There are currently no approved treatments for ASD as a whole, however, treatment regimens have been developed to address specific symptoms related to ASD. Atypical antipsychotics have been evaluated and approved for treating aggressive or self-injurious behavior, severe mood swings, tantrums, and irritability in individuals with ASD. A commonly prescribed, and well-studied, atypical antipsychotic in ASD is risperidone[180]. The primary action of this molecule is serotonin 5-HT_2 receptor blockade. It is also a potent dopamine D_2 receptor antagonist[147]. Selective serotonin reuptake inhibitors (SSRIs) are also often used for treating repetitive behaviors in ASD, and are known to regulate peripheral and central nervous system serotonin levels[110]. SSRIs are effective in treating obsessive compulsive disorder in individuals without a diagnosed ASD[39]. However, current evidence suggests SSRIs, specifically citalopram (or escitalopram) and fluoxetine, are ineffective in treating restrictive repetitive behaviors in individuals with ASD[110, 146]. There has also been a recent push in the medical community to develop treatments that supplement endogenous molecules, like melatonin and oxytocin, shown to have dysregulated production in some ASD patients[22, 137]. Unfortunately, there is insufficient evidence supporting efficacy for most small molecule compounds used to treat ASD symptoms, and a large body of reported adverse events[110, 146, 180]. Further functional characterization of implicated genes and biological pathways are important avenues of research that will hopefully provide results helpful toward more effective personalized treatment of these psychiatric syndromes.

All of the combined research in ASD highlights the incredible complexity of these disorders. It is difficult to identify unifying themes and establish reliable genotype-phenotype relationships. The aim of this project was to overcome issues complicating identification and characterization of genetic factors involved in ASD. We attempted to minimize the effects of phenotypic heterogeneity, locus heterogeneity, epistasis and multiple genes conferring small effects to potentially increase power to detect genetic factors underlying ASD. To progress toward understanding how these significant genetic findings contribute to disease process and identify more effective treatments for ASD, further functional characterization of these associations is necessary. We attempted functional characterization of ASD-associated variation by screening a strongly implicated candidate gene for small molecule effects. This project has the opportunity to broadly impact the biomedical research community by contributing not only to ASD etiology and genetics, but also neurodevelopmental biology and pharmacogenetics.

CHAPTER II

IDENTIFICATION OF GENETICALLY MEANINGFUL PHENOTYPIC SUBGROUPS IN AUTISM SPECTRUM DISORDERS

Introduction

As discussed in Chapter I, genetic factors have a strong influence on risk for Autism Spectrum Disorders (ASD)[160]. However, it has been difficult to identify individual, common genetic factors that replicate across multiple ASD cohorts, or confer large effects on risk[15]. A potential reason is that the wide variability in clinical manifestation can be explained by underlying genetic heterogeneity[32, 70, 89]. Identification of more phenotypically homogeneous subgroups of ASD may help account for this heterogeneity, allowing detection of genetic mechanisms conferring larger risk effects for specific ASD subgroups.

Various attempts have been made to reduce heterogeneity in large-scale genetic studies of ASD. One approach is to separate individuals who meet Diagnostic and Statistical Manual-IV (DSM-IV) criteria for strict Autistic Disorder separately from those meeting only some criteria (i.e. DSM-IV Pervasive Developmental Disorder Not Otherwise Specified [PDD-NOS] or Asperger Disorder)[12, 17, 123, 230]. While this dichotomous categorization of ASD has advanced our knowledge of potential genetic risk factors, via detection of multiple statistically associated and/or linked chromosomal regions, it has still not implicated any genetic variants with large effects[15]. Further, family studies suggest that each of the behavioral domains underlying autism, including social impairment, communication impairment, and repetitive behavior, has separately inherited genetic risk factors that segregate in families[44]. Additionally, the change in criteria between DSM-IV and the new DSM-5 was motivated by the lack of empirical data supporting separate disorders within the autism spectrum, highlighting the need for empirical approaches to identifying subphenotypes within ASD[64, 192].

Previous phenotype-focused studies have emphasized the importance of evaluating multiple sources of behavioral information when attempting to identify behaviorally defined subgroups within ASD[65, 67, 122, 213]. Multivariate statistical methods evaluating multiple sources of behavioral data have been used previously to identify between two and four defined subgroups within the broader classification of ASDs. Categories used to distinguish these previously identified subgroups are severe, moderate and mild ASD, and severe intellectual disabilities[53, 57, 164, 181, 187, 197, 201, 225]. The most consistent findings across these different analyses are subgroups defined as either high- or low-functioning based on the level of symptom severity and some measure of intellectual capability. When age at exam is controlled for, fewer distinct clusters are identified and functional level (as indicated in these studies by nonverbal IQ, Wing Autistic Disorder Interview Checklist, Peabody Picture Vocabulary Tests, and VABS) stands out as a distinct identifier of subgroups[57, 201]. Despite these data, most studies have not evaluated whether or not there are specific genetic contributions to these phenotypic subgroups. One notable exception is a study where subsequent genetic analyses were performed in subgroups defined by cluster analysis[88, 89, 91]. Novel genetic factors were associated with distinct ASD subgroups, providing further support for phenotypic subgroups being genetically meaningful[88]. However, the cluster analysis used to define subgroups was limited to a single source of behavioral information, the ADI-R[91].

Many previous subgrouping efforts also lacked ascertainment of biomarkers or comorbidities commonly seen in ASD. As quantitative traits that are associated with ASD but not required for diagnosis, biochemical or anatomical biomarkers such as elevated whole blood serotonin levels or enlarged head size may improve our ability to identify more genetically homogeneous subgroups[76, 89, 122, 214, 225]. For example, multiple groups have implicated the same chromosomal region, 7q35, and candidate gene, *CNTNAP2*, by refining phenotype definitions to include specific language impairment (SLI) in ASD, which parallels findings in isolated SLI[6, 54, 161, 215, 223]. With the DSM-5, SLI is removed from the ASD criteria and may therefore represent a comorbid diagnosis that is seen in a substantial minority of children with ASD, similar to other comorbid disorders like epilepsy[209].

We hypothesized that subgrouping cases using multiple sources of behavioral and biomarker data would create a more genetically meaningful phenotype definition and increase our power to detect genes influencing risk for ASD. We used novel applications of multivariate statistics to explore behavioral and clinical information from multiple sources.

Methods

Integrate Behavioral and Biomarker Data

We included domain scores from the two main diagnostic instruments, the Autism Diagnostic Interview-Revised (ADI-R)[128] and Autism Diagnostic Observation Schedule (ADOS)[75, 127]. Diagnosis-based studies find the greatest specificity when using both the ADI-R and ADOS in a multidisciplinary assessment process[118]. We also included scores from Vineland Adaptive Behavior Scales (VABS)[193-195] for evaluation of intellectual and adaptive function, an important distinguishing factor in ASD[28, 141]. Ages at exam for all three instruments were included. Finally, we included the quantitative biomarker ‘head circumference’ (HC) as an indicator of either developmental or persistent macrocephaly. While macrocephaly is seen in the minority of adults with ASD, an increased rate of head growth during early childhood is noted in many children with ASD[62, 213].

Multivariate Analyses

We determined the correlation between phenotype traits in the discovery dataset using Spearman's rank correlation coefficients. Since many variables are correlated, and discriminant analyses are extremely sensitive to variable input, we developed a weighting scheme (described below) for input variables based on the correlation structure to ensure that inter-correlated phenotype information did not overly influence the results.

To understand the underlying phenotypic variability in the discovery dataset we performed a Principal Components Analysis (PCA)[87]. This analysis identifies the most important phenotypic traits in the data, simplifies the description of the dataset, and analyzes the structure of the observations and the input variables[2].

To define subgroups of phenotypic expression in the broader diagnostically-defined ASD dataset, we performed agglomerative hierarchical cluster analysis. This clustering method begins with each individual as a separate cluster and aggregates them back together using connectivity-based methods to evaluate the input data, effectively identifying groups of individuals having more similar measures across all input variables[103].

Dataset Demographics

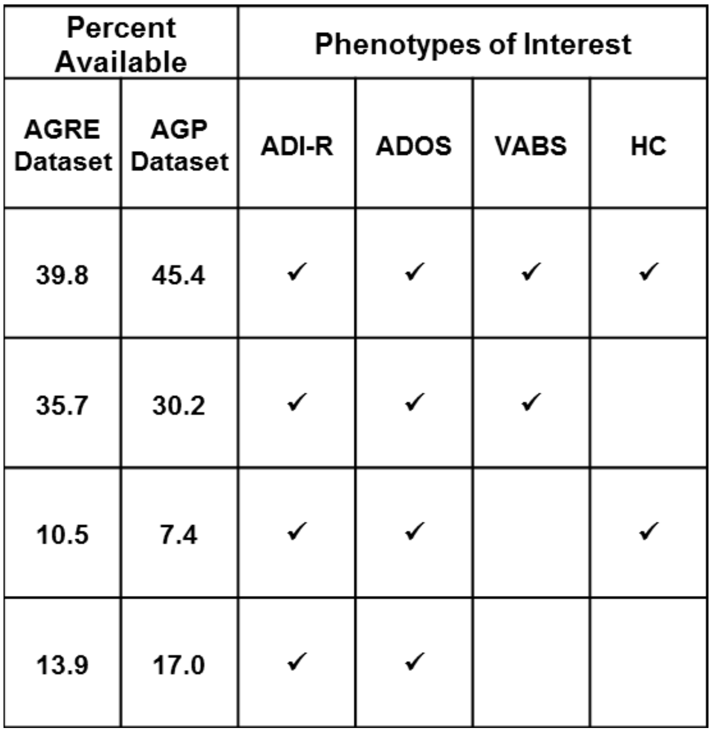
The discovery dataset consists of individuals from the Autism Genetic Resource Exchange (AGRE) family-based study[71]. Individuals not meeting DSM-IV criteria[12] for an Autism Spectrum Disorder diagnosis on both the ADI-R and the ADOS were excluded. We also excluded individuals with potentially non-idiopathic autism (e.g. known neurogenetic disorders, known chromosomal abnormalities, prematurity <35 weeks). The final discovery dataset has 1,261 ASD cases, age at ADI-R 2-21 years old. The genetic ancestry as determined by the software program Structure[165] is 73% European

American (EA), 17.8% Mexican American, 2.7% African American, and 6.5% unknown ethnicity due to missing genome-wide data. This dataset is 80% male and 95% of the cases are from multiplex families.

The dataset we used for replication consists of individuals from the Autism Genome Project (AGP)[92]. This dataset is comprised of 2,563 ASD cases who are not present in the discovery AGRE dataset, meet DSM-IV criteria for a spectrum disorder on both the ADI-R and ADOS, and were 2-21 years old at the time of ADI-R. The genetic ancestry is 64.6% European American, 3% Mexican American, 2% African American, and 30.4% unknown ethnicity due to missing genome-wide data. This dataset is 84% male and 54% of the cases are from multiplex families. The de-identified individual and family IDs for the final datasets are available in Appendix 1.

Phenotype Data Comparisons

We included social, communication, and restricted repetitive behavior (RRB) domain scores from both the ADI-R and ADOS. The communication measure for the ADI-R is divided into verbal and nonverbal scores. Since every person evaluated on the ADI-R receives a nonverbal score but not a verbal score and verbal and nonverbal communication scores are strongly correlated ($\rho=0.86$), we only incorporated the nonverbal scores in our analyses. We also included ‘abnormality of development evident at or before 36 months’ (DevAb) domain scores from the ADI-R. When available, domain standard scores for socialization, communication, daily living skills and motor skills were included from the VABS. Ages at exam for all behavioral tests were also incorporated into analyses. We evaluated head circumference (HC) z-scores taken at one time point. We generated z-scores for available HC measures by standardizing for age and sex using a normal population[170]. We excluded any HC measures taken when individuals were <1 month old. 25%-46% of the VABS and HC data were missing across the datasets; however, the methods we used allow for and are robust to missing data (Table 2.1).



Percent Available		Phenotypes of Interest			
AGRE Dataset	AGP Dataset	ADI-R	ADOS	VABS	HC
39.8	45.4	✓	✓	✓	✓
35.7	30.2	✓	✓	✓	
10.5	7.4	✓	✓		✓
13.9	17.0	✓	✓		

Table 2.1. Availability of Phenotypes. Reported are percentage breakdowns of trait-specific information in both datasets.

Traits included in our analyses represent different types of statistical variables, making direct comparisons difficult. The ADI-R is an interview given by a trained ASD specialist to caregivers of children and adults suspected of having an ASD. It probes for language, social, behavioral and functional abnormalities inconsistent with the individual’s current developmental stage. The ADI-R interview generates scores in each of three content areas: communication and language, social interaction, and restricted, repetitive behaviors. Item scores are measured on a finite ordinal scale. Increased scores indicate more severe abnormalities reported for the evaluated behaviors[126, 128]. Domain scores are calculated for all items assessing the behavioral characteristics relevant to ASD (social, communication, and restricted repetitive behaviors) and represent the sum of relevant item scores.

ADOS is a semi-structured assessment of communication, social interaction and play, or imaginative use of materials, for individuals suspected of having autism or other pervasive developmental disorders[127]. Behavioral items relevant to ASD are scored on finite, ordinal scales, higher scores on these items indicate increased severity for abnormalities in the evaluated behavior[75]. Domain scores are calculated as described above for the ADI-R. ADOS domain scores were modified prior to percentile rank calculations to be comparable across the four possible modules by reducing raw ordinal values to that of the module with the smallest scale for each domain. For example, for ADOS modules 1, 3, and 4 communication is scored on an ordinal scale from 0-6, while for ADOS module 2 this measure is only scored on a scale from 0-4. Therefore, communication domain scores from modules 1, 3, and 4 were reduced to a scale of 0-4 to make these scores more comparable to module 2.

VABS focuses on social skills and is the measurement of adaptive behaviors, including the ability to cope with environmental changes, to learn new everyday skills and to demonstrate independence. This scale also yields composite and domain scores, however measured on a finite, continuous scale[195]. Increased scores on VABS measures indicate decreased severity for expression of evaluated traits. VABS data were ranked inversely to account for the inverse relationship of these severity scores when compared to the other diagnostic methods used in analyses.

Head circumference z-scores and ages at exams represent continuous variables measured on an infinite scale. To allow more comparable measures, we chose to transform variables into Hazen percentile ranks using Stata 11.2[84, 200].

We determined the correlation structure across all these variables by calculating pairwise Spearman's rank correlation coefficients (ρ) (Stata 11.2) using all available percentile rank data.

Item-Level & Domain Score Comparisons: ADI-R & ADOS

We chose to use domain scores, as opposed to item-level scores, from all evaluated behavioral instruments since these scores effectively cover information relevant to primary phenotype characteristics in ASD, and to minimize the potential for overfitting in our cluster analyses. Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit will generally have poor predictive performance, as it can exaggerate minor fluctuations in the data[207]. However, item level scores provide potentially genetically-relevant phenotypic information related to endophenotypes (i.e. savant skills)[88, 89]. To determine what information relevant to item-level scores were not covered by domain scores included in our analyses, we calculated percentile ranks for item-levels score from both instruments and determined the correlation across item scores and domain scores assessed on the same instrument.

Principal Components Analysis (PCA)

PCA was performed on percentile ranked data using the ‘FactoMineR’ package in R[119]. Variable weights were incorporated into PCA using the correlation structure observed in the dataset. We chose a threshold for independence at $\rho<0.50$. If a variable was correlated with another variable at $\rho\geq0.50$ those variables were weighted to allow for only a partial variable contribution to PCA. Social and communication domain scores from the ADI-R were weighted such that these two scores together contributed one total variable weight in analysis. ADI-R RRB measures did not meet our threshold for correlation with any other variable and were therefore independent of other variables included in analysis. ‘Developmental abnormality evident prior to 36 months’ domain scores were also given one total variable weight. For the ADOS, we weighted social and communication domain scores together as one total variable contribution. The ADOS

RRB domain scores were weighted as an independent contribution. For the VABS, all of the domain standard scores were weighted as one total variable contribution. It is notable that the strongest correlations observed for the motor skills domain standard scores are with the communication domain standard scores at ρ=0.49, which did not quite meet our threshold for non-independence. However, the correlations observed by the VABS developers for the motor skills domain standard scores indicated dependence on the communication domain standard score (ρ=0.56-0.61)[193]. As such, we chose to incorporate only a partial weight for motor skills domain standard scores in our analyses. Head circumferences were given one total variable weight. Ages at exam for ADI-R, ADOS and VABS were weighted such that these three variables contributed one total variable weight. The cumulative number of variables incorporated into PCA using this weighting scheme equaled eight variables. We allowed up to 20 PCA dimensions to be retained in the results.

Optimal Clustering Method and Dataset Partitions

Dissimilarity matrices were calculated using the Gower dissimilarity measure from the 'FD' package in R, with variables weighted according to the weighting scheme described above for PCA[115, 116]. Seven different clustering methods were evaluated for internal validity while partitioning the dissimilarity matrix into anywhere from two to 15 clusters using the 'clValid' package in R[31]. Evaluated clustering methods were kmeans, agglomerative hierarchical, model-based, partitioning around medoids, divisive hierarchical, self-organizing tree algorithm, and clustering large applications.

To evaluate cluster validity, clValid calculates the Connectivity (an indication of the degree of connectedness of the clusters), Dunn index (a ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance) and Silhouette Width (the overall average of the average distance between each observation and others in the same cluster compared to different clusters). To evaluate cluster stability, clValid calculates the Average Proportion of Non-overlap (average proportion of observations not placed in the same cluster when variables are removed), Average Distance (average distance between observations placed in the same cluster when variables are removed), Average Distance between Means (average distance between cluster centers for observations placed in the same cluster when variables are removed) and Figure of Merit (average intra-cluster variance of the removed variable, where the clustering is based on the remaining variables)[31]. Sensitivity analysis was performed by removing one variable, reapplying weights to account for the missing variable, calculating a Gower dissimilarity matrix, clustering the data and calculating the above mentioned stability scores. This was done for each variable.

Clustering and Cluster Validation

Dissimilarity matrices were calculated as described above and variables were weighted according to the weighting scheme described above for PCA[115, 116]. The final agglomerative hierarchical clustering was performed on the Gower dissimilarity matrix using the ‘cluster’ package in R[135]. The agglomerative coefficient was calculated for the final clustering of the data. This represents a measure of all the individual dissimilarities calculated across the dataset and is an indication of the clustering structure identified[103]. This coefficient is measured on a scale from zero to one, zero indicating no clustering structure and one indicating complete structure.

Validity of the final clusters was determined by permuting phenotype data across individuals, clustering the permuted data and calculating the Adjusted Hubert-Arabie Rand index (AHARI) to compare clustering of the real data to the permuted data[94]. This was done for 1,000 data permutations and the AHARIs were averaged. The permutations were accomplished by writing a function in R and the AHARI statistic was calculated using a command from the ‘mclust’ package[63]. Sensitivity analyses were performed using the ‘clValid’ package in R, with slight modifications; weights were reapplied to account for variables removed and a Gower dissimilarity matrix was calculated prior to clustering[31]. Kruskal-Wallis tests were performed in STATA 11.2 on untransformed scores to determine the distributional variation of scores between main clusters and across subclusters.

Genetic Contribution to Cluster Assignment

Intra-cluster family structure was evaluated by calculating the odds of individuals being assigned to the same cluster given a familial relationship. We generated a 2X2 contingency table and calculated an odds ratio via the chi-square statistic. ‘Case’ status was defined as a full sibling relationship and ‘exposure’ was defined as assignment to the same phenotype cluster. Each individual was manually scored for the number of full sibling relationships in the dataset. Since there are substantially more unrelated individuals than related in the datasets, we randomly sampled groups of unrelated individuals representing the same number of available familial relationships. We calculated an odds ratio for related ‘cases’ and each randomly sampled unrelated ‘control’ group. This was done 10 times. The reported odds ratios represent the range for these calculations. We estimated genetic relationships using Single Nucleotide Polymorphism (SNP) markers previously genotyped in our datasets. Markers were pruned using genotyped founders based on linkage disequilibrium. We set an r^2 threshold of 0.16, within a 500 SNP window, sliding 5 SNPs at a time. We subsequently created a pedigree file of cases in our cluster dataset. Wright’s F-statistic (Fst) was then calculated using PLATO[77]. We grouped individuals into subpopulations based on cluster assignment. For each genetic marker, the correlation between individuals drawn from the subpopulation relative to the total population was determined. We then took the average Fst calculated across the informative autosomal markers.

Results

Discovery Dataset (AGRE)

Correlation Among Variables

The correlation structure indicates diverse relationships among phenotype variables in the AGRE dataset (Fig. 2.1; Table 2.2). Social and communication scores measured on the same instrument are positively correlated ($\rho_{ADI\text{-}R}$=0.62, ρ_{ADOS}=0.57, ρ_{VABS}=0.80), while restricted and repetitive behavior scores are not strongly correlated with social and communication scores assessed on the same instrument ($\rho_{ADI\text{-}R}$=0.07, 0.17; ρ_{ADOS}=0.18, 0.35). When comparing scores evaluating the same behavioral characteristic between the ADOS and ADI-R instruments, there is minimal correlation, especially with regard to RRB scores (ρ_{Social}=0.37, $\rho_{Communication}$=0.31, ρ_{RRB}=0.04). The strongest variable correlations across the ADI-R, ADOS and VABS are positive correlations between the social and communication scores from the ADI-R and VABS (ρ=0.45). The strongest correlation for the ‘developmental abnormality evident prior to 36 months’ scores from the ADI-R are a positive relationship with ADI-R social and communication domain scores (ρ_{Social}=0.31, $\rho_{Communication}$=0.29). Head circumferences are not strongly correlated with any of the behavioral measures. The strongest correlation for HCs is a positive correlation with VABS social and daily living skills domain standard scores (ρ=0.14). As expected, ages at exam are strongly correlated across the ADI-R, ADOS and VABS (ρ=0.84-0.94).

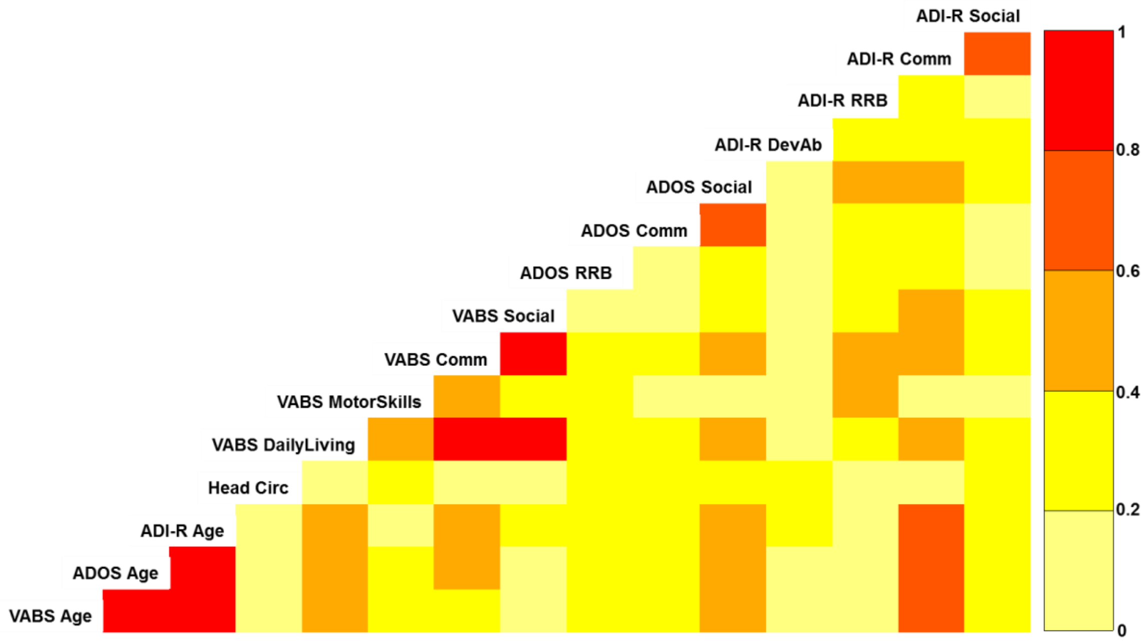


Figure 2.1. Variable Correlation Structure in Discovery Dataset. Plot of Spearman's correlation coefficients used in variable weighting scheme for PCA and clustering.

Variable	ADI-R Social	ADI-R Comm	ADI-R RRB	ADI-R DevAb	ADOS Social	ADOS Comm	ADOS RRB	VABS Social	VABS Comm	VABS MS	VABS DL	HC	ADI-R Age	ADOS Age	VABS Age
ADI-R Social	1.00														
ADI-R Comm	0.62	1.00													
ADI-R RRB	0.17	0.07	1.00												
ADI-R DevAb	0.31	0.29	-0.02	1.00											
ADOS Social	0.37	0.39	-0.04	0.26	1.00										
ADOS Comm	0.27	0.31	0.04	0.23	0.57	1.00									
ADOS RRB	0.21	0.31	0.04	0.19	0.35	0.18	1.00								
VABS Social	0.45	0.42	0.09	0.18	0.41	0.24	0.29	1.00							
VABS Comm	0.40	0.45	0.00	0.27	0.49	0.28	0.34	0.80	1.00						
VABS MS	0.22	0.33	0.01	0.27	0.33	0.22	0.34	0.37	0.49	1.00					
VABS DL	0.33	0.39	0.07	0.22	0.40	0.27	0.31	0.79	0.80	0.49	1.00				
HC	0.09	0.13	-0.11	-0.03	0.11	0.04	0.10	0.14	0.13	0.05	0.14	1.00			
ADI-R Age	0.26	0.00	0.17	-0.13	-0.12	-0.12	-0.20	0.22	0.05	-0.38	-0.02	0.11	1.00		
ADOS Age	0.16	-0.03	0.14	-0.16	-0.10	-0.11	-0.16	0.31	0.13	-0.38	0.08	0.14	0.89	1.00	
VABS Age	0.15	-0.02	0.10	-0.17	-0.09	-0.11	-0.12	0.35	0.16	-0.39	0.11	0.14	0.84	0.94	1.00

Table 2.2. Spearman's Correlation Coefficients. Spearman's rho correlations calculated in AGRE discovery dataset. Comm=Communication Domain Scores; RRB=Restricted, repetitive behaviors; DevAb=Abnormality of Development evident ≤36 months; MS=MotorSkills; DL=Daily Living; HC=head circumferences.

Item-Level & Domain Score Comparisons: ADI-R & ADOS

Spearman’s correlation coefficients indicate that for the ADI-R, the domain scores we included in analyses do not provide information relative to presentation of savant skills, acts of aggression, or hyperactivity (Fig 2.2a). Domain scores from the ADOS do not provide information relative to speech abnormalities associated with ASD, anxiety, aggressive tendencies, or hyperactivity (Fig 2.2b).

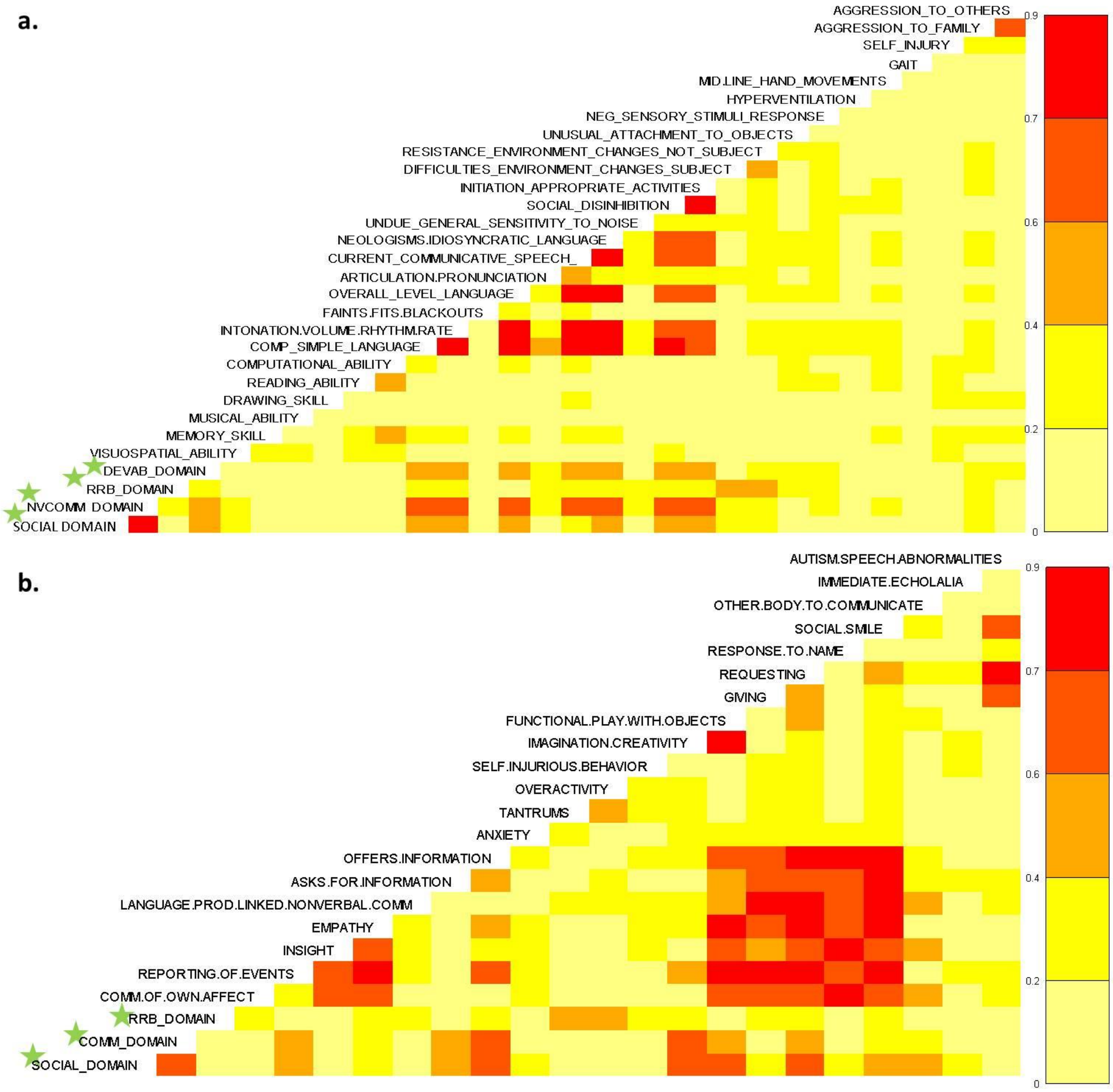


Figure 2.2. Correlation Across Domain and Item Scores. Plot of Spearman’s correlation coefficients showing correlation across domain scores used as variable input (indicated by stars) and item-level information not included in domain score calculations for **a.** ADI-R and **b.** ADOS.

Principal Components Analysis

PCA identifies15 components comprising the data, with 53% of the phenotypic variance being explained by the first three components and the remainder of the variance being explained in increasingly smaller increments from components four to 15 (Fig. 2.3). Principal component (PC) one defines 25% of the phenotypic variance in the discovery dataset. Although most input variables contribute to the phenotypic variance defined in PC1, the two variables with the strongest contributions are ADOS RRB scores and ADI-R DevAb scores (Table 2.3). HC, ADI-R RRB scores and ages at exam do not have strong contributions to PC1. However, these variables explain the majority of the phenotypic variance defined by PC2 and PC3. These two components combined explain another 29% of the phenotypic variance in the discovery dataset (PC2≈15%, PC3≈14%). PC4 defines another 11% of the phenotype variation in the dataset. Similar to PC1, the two variables contributing most to the phenotypic variance defined in PC4 are ADI-R DevAb scores and ADOS RRB scores. However, unlike PC1, the next strongest contributors are RRB scores from the ADI-R. Social and communication scores from the ADI-R and ADOS, and scores from the VABS have much smaller contributions to PC4 than to PC1. PC5 defines 9% of the variance in the data and has strong contributions from HCs, DevAb scores and RRB scores from the ADI-R. PC6 defines another 7.5% of the variance in the dataset with ADOS RRB and communication scores contributing to over half of this defined variance. PC7 defines another 5.5% of the phenotypic variance; its strongest contributors are measures from the VABS and ADOS communication scores. PC8 defines another 5% with the strongest contributors being ADI-R social and communication scores closely followed by these same scores from the ADOS. The combined phenotypic variance explained in the AGRE dataset by the first 8 principal components is 91.5%. The remaining principal components, PC 9-15, each define very small portions of the phenotypic variance observed in the data (0.35%-2.5%) and combined explain the remaining 8.5% of phenotypic variance in the dataset. The variables contributing the most to these final seven PCs are further outlined in Table 2.3.

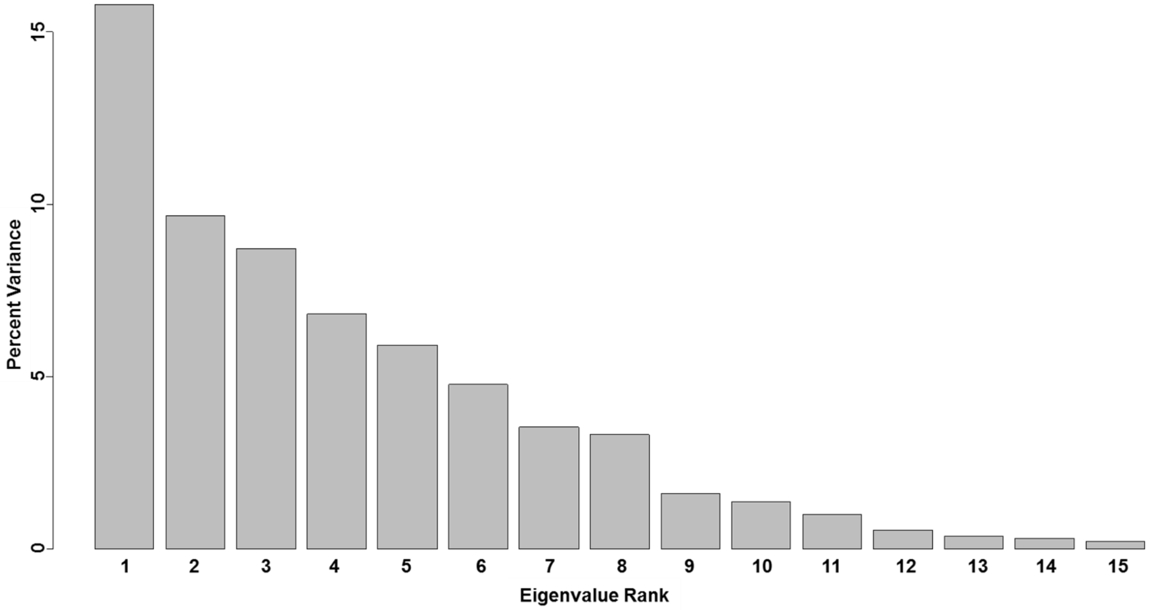


Figure 2.3. Phenotype Variance Explained by Principal Components. Plotted are the percentages of phenotypic variance explained, based on eigenvalues, by each Principal Component defined in the AGRE dataset.

Phenotype Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10	Comp11	Comp12	Comp13	Comp14	Comp15
ADI-R Social	9.47	5.90	0.25	3.56	0.22	0.60	2.38	***26.14***	5.61	***36.13***	8.03	0.45	0.03	0.67	0.57
ADI-R Comm	11.62	1.12	0.00	0.60	0.39	2.27	0.33	***36.35***	10.11	***34.17***	2.96	0.03	0.03	0.01	0.00
ADI-R RRB	0.32	***29.47***	***32.83***	13.33	***20.98***	0.83	0.32	1.02	0.17	0.65	0.01	0.03	0.05	0.01	0.00
ADI-R DevAb	***19.06***	4.56	2.91	***30.17***	***21.00***	18.11	0.01	3.74	0.00	0.38	0.02	0.00	0.02	0.02	0.00
ADOS Social	11.80	0.04	0.27	0.03	2.24	13.07	5.06	6.78	***50.51***	8.47	1.18	0.01	0.49	0.00	0.06
ADOS Comm	7.30	0.04	0.05	0.36	0.13	***24.48***	15.66	15.23	***31.32***	5.30	0.01	0.01	0.11	0.00	0.00
ADOS RRB	***22.07***	0.97	0.07	***37.16***	6.67	***27.32***	4.40	0.33	0.44	0.47	0.00	0.11	0.00	0.01	0.01
VABS Social	3.80	3.09	0.20	0.39	3.81	0.11	14.46	1.38	0.19	0.02	5.42	1.50	0.00	***65.30***	0.35
VABS Comm	4.90	0.85	0.23	0.27	2.87	0.31	16.71	1.48	0.01	0.07	0.96	2.41	***57.33***	11.59	0.02
VABS MotorSkills	3.25	0.91	0.02	0.53	0.02	1.33	15.29	0.15	0.00	9.55	***51.96***	16.18	0.63	0.20	0.00
VABS DailyLiving	4.13	0.81	0.12	0.03	1.82	0.62	***21.22***	1.80	0.99	0.01	3.65	11.97	***39.31***	11.63	1.90
HC	1.18	6.12	***61.59***	2.92	***28.00***	0.08	0.00	0.04	0.00	0.01	0.03	0.00	0.01	0.00	0.00
ADI-R Age	0.45	***16.14***	0.19	4.34	2.44	3.15	2.62	0.13	0.02	1.57	14.16	12.91	0.59	2.60	***38.71***
ADOS Age	0.42	***16.24***	0.51	3.47	3.95	3.58	1.00	2.26	0.07	0.00	8.42	4.11	0.75	0.22	***55.01***
VABS Age	0.26	***13.74***	0.77	2.85	5.48	4.15	0.54	3.18	0.58	3.21	3.21	***50.28***	0.65	7.75	3.38

Table 2.3. Variable Contributions to Principal Components of AGRE Dataset. Variables contributing the most to the observed variance explained by each component are indicated in bold italics.

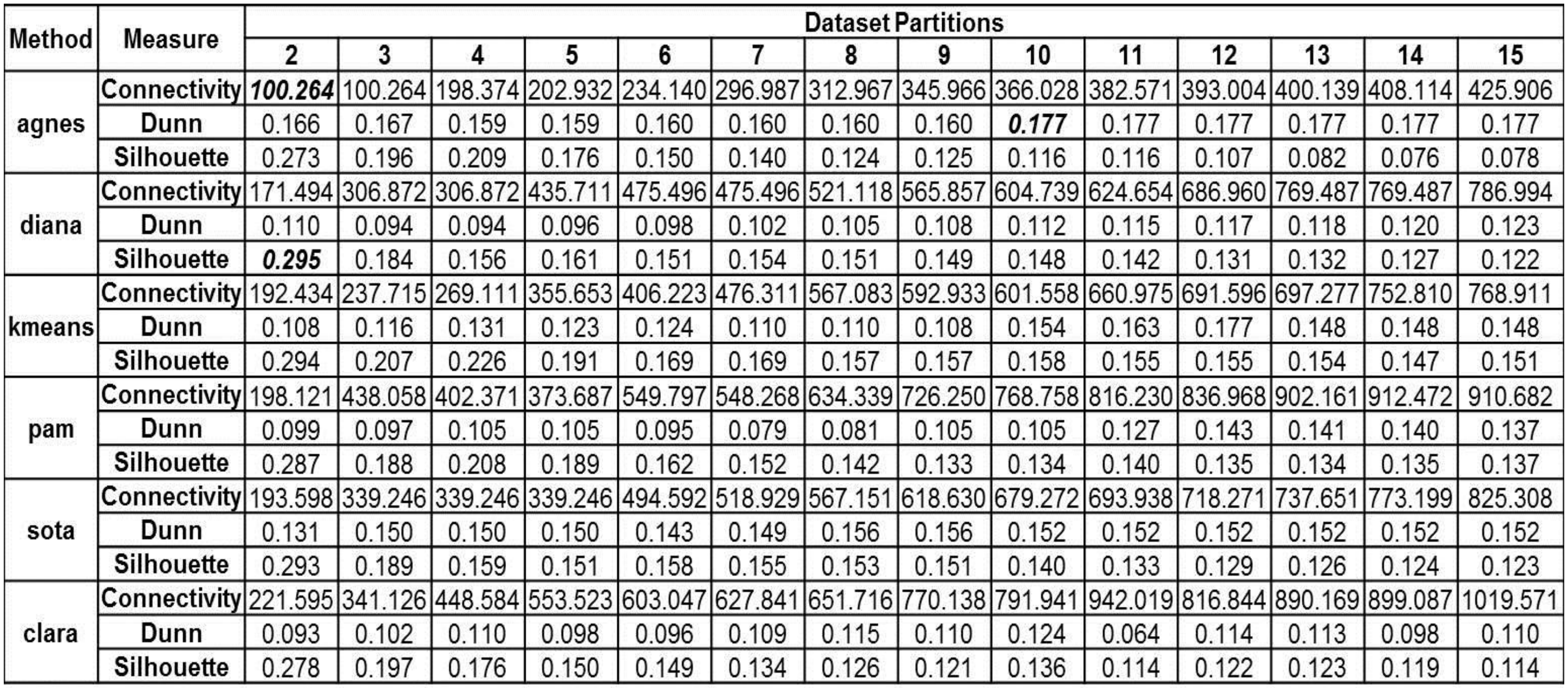
Clustering

PCA helped to define the underlying phenotypic variability in the dataset and identify the most important classifying variables, but did not clarify the phenotypic nature of each subgroup of cases. Unsupervised clustering was therefore performed to define ASD subgroups and obtain a broader sense of the phenotype characteristics of these subgroups. The overall best validity scores were calculated when using agglomerative hierarchical clustering to group the AGRE dataset into two clusters. The next best validity scores were calculated when using agglomerative hierarchical clustering to subgroup the dataset into 10 subclusters (Table 2.4).

Following agglomerative hierarchical clustering, we grouped the data into the most valid partition (i.e. two major clusters), one cluster with 443 cases and one cluster with 818 cases (Fig. 2.4). The agglomerative coefficient calculated for clustering of the AGRE dataset is 0.78, evidence that a strong clustering structure was identified. We evaluated phenotype variable distributions between the two main clusters. Kruskal-Wallis tests show that all variable distributions, except ADI-R RRB and HC, are significantly different (p<0.0001) between these clusters (Table 2.5). Examination of the summary statistics for phenotype variables by cluster show that individuals with scores indicating more severe measures for most variables are placed into the larger cluster, referred to as ‘more severe’, when compared to the smaller cluster, referred to as ‘less severe’ (Table 2.6). The two main clusters could then be grouped into 10 subclusters; the ‘more severe’ main cluster grouped into six subclusters and the ‘less severe’ main cluster grouped into four subclusters. Phenotype variable distributions were then evaluated across the 10 subclusters. Kruskal-Wallis tests show that the previously non-significant ADI-R RRBs and HC are very different (p<0.0001) across the 10 subclusters. HC distributions are statistically different across the four subclusters comprising the ‘less severe’ main cluster (p=0.0034) and the six subclusters comprising the ‘more severe’ main cluster (p<0.0001). ADI-R RRB score distributions are also statistically different between the four subclusters comprising the ‘less severe’ main cluster (p<0.0001) and the six subclusters comprising the ‘more severe’ main cluster (p<0.0001). The average Adjusted Hubert-Arabie Rand index (AHARI) calculated over 1,000 data permutations shows that partitioning of real data for the discovery dataset is significantly different than partitioning permuted datasets (AHARI=$-6.14x10^{-5}$).

Sensitivity analyses show that ADI-R DevAb scores have the overall largest effect on main cluster stability. Communication scores from the ADOS and social, communication and daily living domain standard scores from the VABS appear to have the least effect on main cluster stability. The remaining input variables have similar and modest effects on main cluster stability. Regarding the subclusters, with the exception of the DevAB scores from the ADI-R, removal of any other input variable has similar and minor effects on subcluster stability (Table 2.7).

Familial relationships are significantly associated with assignment to the two main phenotype clusters (OR≈1.38-1.42, p<0.00001). Wright’s F-statistic indicates that genotype frequencies are more similar within clusters than in the entire unclustered dataset (Average Fst≈0.17) (Table 2.8).



Method	Measure	Dataset Partitions													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
agnes	Connectivity	***100.264***	100.264	198.374	202.932	234.140	296.987	312.967	345.966	366.028	382.571	393.004	400.139	408.114	425.906
	Dunn	0.166	0.167	0.159	0.159	0.160	0.160	0.160	0.160	***0.177***	0.177	0.177	0.177	0.177	0.177
	Silhouette	0.273	0.196	0.209	0.176	0.150	0.140	0.124	0.125	0.116	0.116	0.107	0.082	0.076	0.078
diana	Connectivity	171.494	306.872	306.872	435.711	475.496	475.496	521.118	565.857	604.739	624.654	686.960	769.487	769.487	786.994
	Dunn	0.110	0.094	0.094	0.096	0.098	0.102	0.105	0.108	0.112	0.115	0.117	0.118	0.120	0.123
	Silhouette	***0.295***	0.184	0.156	0.161	0.151	0.154	0.151	0.149	0.148	0.142	0.131	0.132	0.127	0.122
kmeans	Connectivity	192.434	237.715	269.111	355.653	406.223	476.311	567.083	592.933	601.558	660.975	691.596	697.277	752.810	768.911
	Dunn	0.108	0.116	0.131	0.123	0.124	0.110	0.110	0.108	0.154	0.163	0.177	0.148	0.148	0.148
	Silhouette	0.294	0.207	0.226	0.191	0.169	0.169	0.157	0.157	0.158	0.155	0.155	0.154	0.147	0.151
pam	Connectivity	198.121	438.058	402.371	373.687	549.797	548.268	634.339	726.250	768.758	816.230	836.968	902.161	912.472	910.682
	Dunn	0.099	0.097	0.105	0.105	0.095	0.079	0.081	0.105	0.105	0.127	0.143	0.141	0.140	0.137
	Silhouette	0.287	0.188	0.208	0.189	0.162	0.152	0.142	0.133	0.134	0.140	0.135	0.134	0.135	0.137
sota	Connectivity	193.598	339.246	339.246	339.246	494.592	518.929	567.151	618.630	679.272	693.938	718.271	737.651	773.199	825.308
	Dunn	0.131	0.150	0.150	0.150	0.143	0.149	0.156	0.156	0.152	0.152	0.152	0.152	0.152	0.152
	Silhouette	0.293	0.189	0.159	0.151	0.158	0.155	0.153	0.151	0.140	0.133	0.129	0.126	0.124	0.123
clara	Connectivity	221.595	341.126	448.584	553.523	603.047	627.841	651.716	770.138	791.941	942.019	816.844	890.169	899.087	1019.571
	Dunn	0.093	0.102	0.110	0.098	0.096	0.109	0.115	0.110	0.124	0.064	0.114	0.113	0.098	0.110
	Silhouette	0.278	0.197	0.176	0.150	0.149	0.134	0.126	0.121	0.136	0.114	0.122	0.123	0.119	0.114

Table 2.4. Validity Scores for Clustering. The most valid number of partitions, and cluster method are indicated in bold italics. agnes=agglomerative hierarchical, diana=divisive hierarchical, pam=partitioning around medoids, sota=self-organizing tree algorithm, clara=clustering large applications. Dunn Index – smallest inter-cluster distance / largest intra-cluster distance, should be maximized (scale=0, ∞). Connectivity - to what extent are individuals placed in the same cluster as the most similar individuals, should be minimized (scale=0, ∞). Silhouette Width - overall average of average distance between individual and others in same cluster compared to different cluster (-1, 1); well-clustered=1.

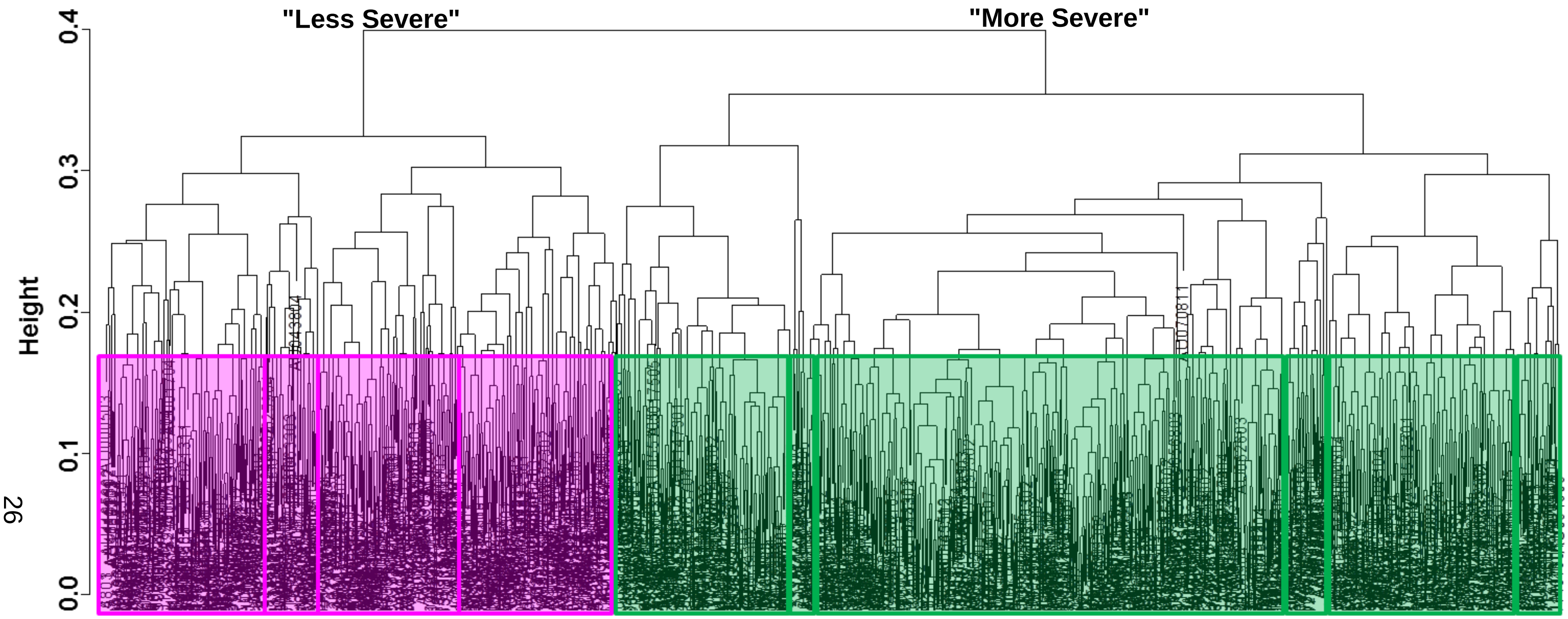
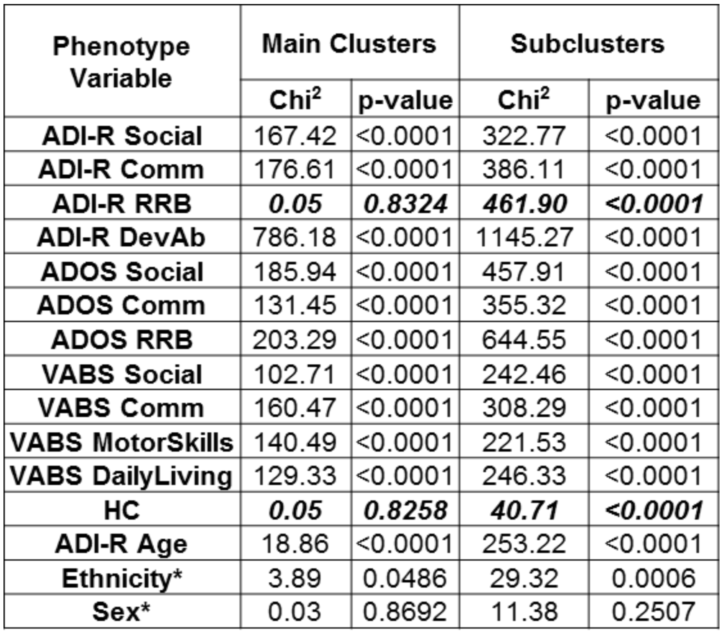
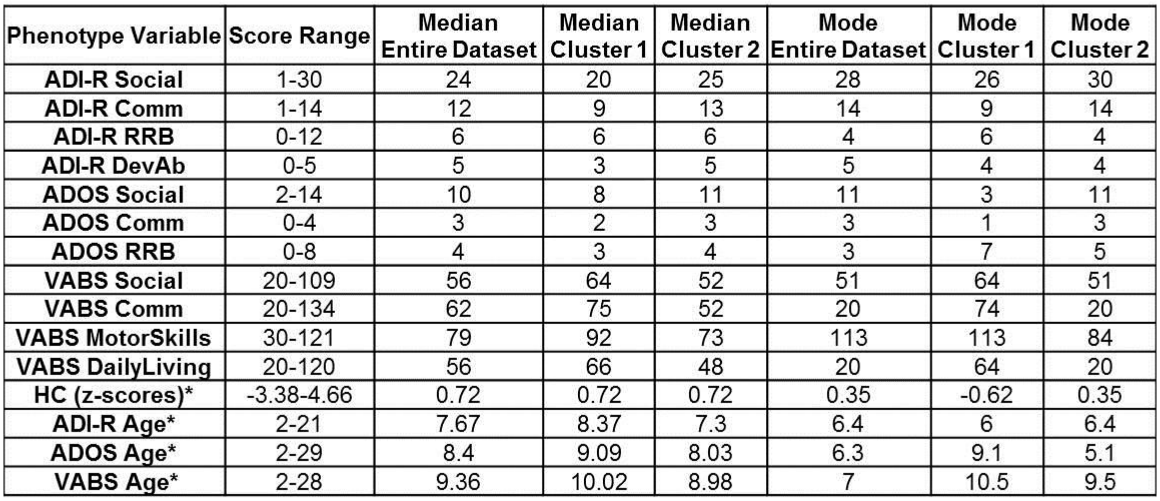


Figure 2.4. Agglomerative Clustering AGRE Dataset. Height indicates distance between merging clusters at successive stages of clustering and is related to dissimilarities among clusters. The main cluster highlighted in green represents individuals with more severe ASD phenotypes compared to individuals assigned to the main cluster highlighted pink. Subclusters are indicated with corresponding boxes.



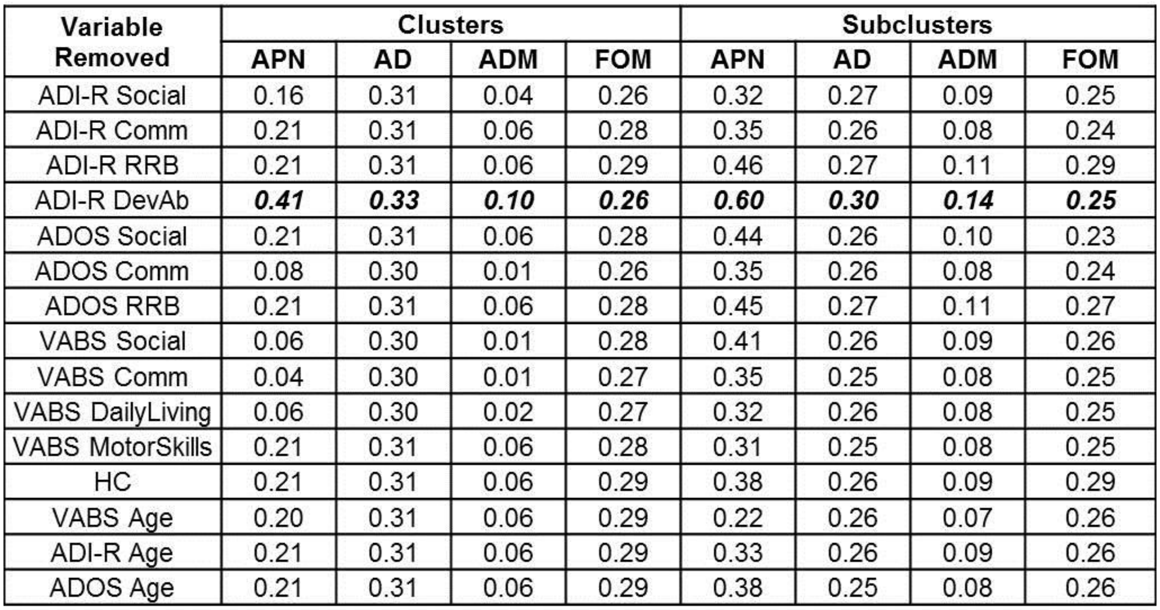
Phenotype Variable	Main Clusters		Subclusters	
	Chi^2	p-value	Chi^2	p-value
ADI-R Social	167.42	<0.0001	322.77	<0.0001
ADI-R Comm	176.61	<0.0001	386.11	<0.0001
ADI-R RRB	***0.05***	***0.8324***	***461.90***	***<0.0001***
ADI-R DevAb	786.18	<0.0001	1145.27	<0.0001
ADOS Social	185.94	<0.0001	457.91	<0.0001
ADOS Comm	131.45	<0.0001	355.32	<0.0001
ADOS RRB	203.29	<0.0001	644.55	<0.0001
VABS Social	102.71	<0.0001	242.46	<0.0001
VABS Comm	160.47	<0.0001	308.29	<0.0001
VABS MotorSkills	140.49	<0.0001	221.53	<0.0001
VABS DailyLiving	129.33	<0.0001	246.33	<0.0001
HC	***0.05***	***0.8258***	***40.71***	***<0.0001***
ADI-R Age	18.86	<0.0001	253.22	<0.0001
Ethnicity*	3.89	0.0486	29.32	0.0006
Sex*	0.03	0.8692	11.38	0.2507

Table 2.5. Cluster Differences in the AGRE Dataset. Kruskal Wallis comparisons of variable distributions between the two main clusters and across the ten subclusters. All input variable distributions, except ADI-R RRB and HC, are significantly different between the main clusters. ADI-R and HC distributions are significantly different across subclusters. Asterisks indicate information not used as input variable.



Phenotype Variable	Score Range	Median Entire Dataset	Median Cluster 1	Median Cluster 2	Mode Entire Dataset	Mode Cluster 1	Mode Cluster 2
ADI-R Social	1-30	24	20	25	28	26	30
ADI-R Comm	1-14	12	9	13	14	9	14
ADI-R RRB	0-12	6	6	6	4	6	4
ADI-R DevAb	0-5	5	3	5	5	4	4
ADOS Social	2-14	10	8	11	11	3	11
ADOS Comm	0-4	3	2	3	3	1	3
ADOS RRB	0-8	4	3	4	3	7	5
VABS Social	20-109	56	64	52	51	64	51
VABS Comm	20-134	62	75	52	20	74	20
VABS MotorSkills	30-121	79	92	73	113	113	84
VABS DailyLiving	20-120	56	66	48	20	64	20
HC (z-scores)*	-3.38-4.66	0.72	0.72	0.72	0.35	-0.62	0.35
ADI-R Age*	2-21	7.67	8.37	7.3	6.4	6	6.4
ADOS Age*	2-29	8.4	9.09	8.03	6.3	9.1	5.1
VABS Age*	2-28	9.36	10.02	8.98	7	10.5	9.5

Table 2.6. Summary Statistics for Unclustered vs Clustered AGRE Datasets. Reported are medians and modes observed in the unclustered dataset compared to the two main clusters. Continuous variables are starred to indicate that the mean is reported in place of the median. Cases with scores indicating increased ASD severity preferentially cluster into the second, larger cluster. Age is reported in years.



Variable Removed	Clusters				Subclusters			
	APN	AD	ADM	FOM	APN	AD	ADM	FOM
ADI-R Social	0.16	0.31	0.04	0.26	0.32	0.27	0.09	0.25
ADI-R Comm	0.21	0.31	0.06	0.28	0.35	0.26	0.08	0.24
ADI-R RRB	0.21	0.31	0.06	0.29	0.46	0.27	0.11	0.29
ADI-R DevAb	***0.41***	***0.33***	***0.10***	***0.26***	***0.60***	***0.30***	***0.14***	***0.25***
ADOS Social	0.21	0.31	0.06	0.28	0.44	0.26	0.10	0.23
ADOS Comm	0.08	0.30	0.01	0.26	0.35	0.26	0.08	0.24
ADOS RRB	0.21	0.31	0.06	0.28	0.45	0.27	0.11	0.27
VABS Social	0.06	0.30	0.01	0.28	0.41	0.26	0.09	0.26
VABS Comm	0.04	0.30	0.01	0.27	0.35	0.25	0.08	0.25
VABS DailyLiving	0.06	0.30	0.02	0.27	0.32	0.26	0.08	0.25
VABS MotorSkills	0.21	0.31	0.06	0.28	0.31	0.25	0.08	0.25
HC	0.21	0.31	0.06	0.29	0.38	0.26	0.09	0.29
VABS Age	0.20	0.31	0.06	0.29	0.22	0.26	0.07	0.26
ADI-R Age	0.21	0.31	0.06	0.29	0.33	0.26	0.09	0.26
ADOS Age	0.21	0.31	0.06	0.29	0.38	0.25	0.08	0.26

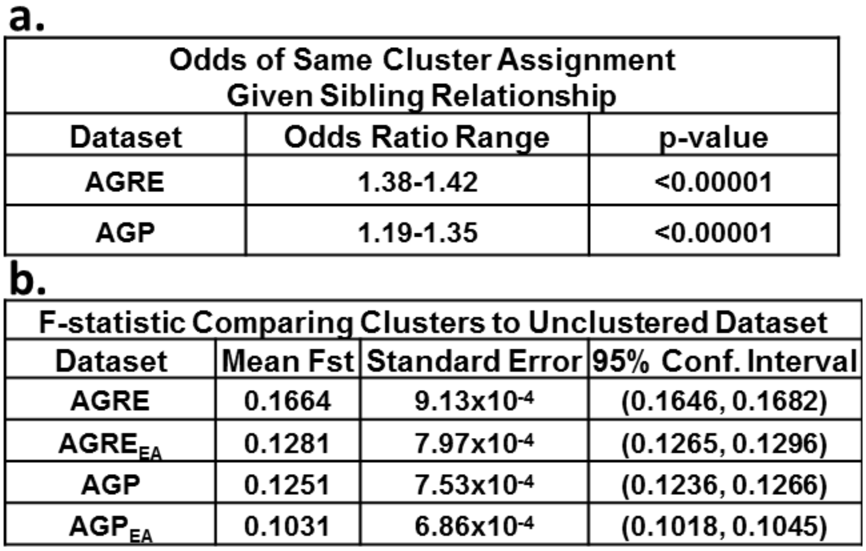
Table 2.7. Sensitivity Analyses. Reported are results from sensitivity analyses. For the stability measures calculated, smaller values indicate more stable cluster results. Statistics evaluating cluster stability upon removal of each variable are: APN=Average proportion of nonoverlap or number of individuals not placed in same cluster when variable is removed (scale=0,1); AD= Average distance between individuals placed in same cluster when variable is removed (scale=0, ∞); ADM=Average distance between means between cluster centers for individuals placed in same cluster when variable is removed (scale=0, ∞); FOM=Figure of merit or average intra-cluster variance of the removed variable where clustering is based on remaining variables (scale=0, ∞).

Replication Dataset (AGP)

We tested for replication in the independent, non-overlapping Autism Genome Project dataset. We see a similar correlation structure among AGP dataset phenotype input variables as in the AGRE dataset (Table 2.9). Using the same correlation threshold (ρ≥0.50), we incorporated the same eight variable weighting scheme in subsequent PCA and clustering analyses. To define the phenotypic variance, PCA again identified 15 components. Most input variables contribute similarly to phenotypic variance explained in PC1, with the exception of HC, ADI-R RRBs, ages at exams and VABS motor skills having little contribution. HC, ADI-R RRB and ages explain the majority of the phenotypic variance defined by PC2 and PC3. Combined, PCs 1-3 define ~50% of the phenotypic variance in the data. Further details on variable contributions to all 15 data components are outlined in Table 2.10. Again, the optimal clustering method determined to group the AGP dataset was determined to be agglomerative hierarchical (Table 2.11). This method validly grouped the AGP dataset into two main clusters and 15 subclusters (Fig. 2.5). Kruskal-Wallis tests show that most input variable distributions are significantly different between the two main clusters ($p<0.0001$), with the exception of HC (Table 2.12). However, the distributions of HC are significantly different across the 15 subclusters ($p=0.0020$). HCs are statistically different between the six subclusters comprising the 'less severe' main cluster ($p=0.0007$) but not the nine subclusters comprising the 'more severe' main cluster ($p=0.37$). Cases with increased severity measures for most variables tended to group into the larger main cluster (n=1,527) compared to the smaller main cluster (n=1,036) containing cases with generally less severe scores for the majority of variables (Table 2.13). The agglomerative coefficient calculated for clustering of the AGP dataset is 0.79, indicating strong hierarchical clustering structure. The AHARI statistic shows that clustering of the real phenotype data is significantly different than permuted datasets (AHARI=-4.10×10^{-6}).

Sensitivity analyses again show that ADI-R DevAb scores have the overall largest effect on main cluster stability. The remaining input variables have similar and modest effects on main cluster stability. Regarding the subclusters, removal of any input variable has similar effects on subcluster stability (Table 2.14).

In the AGP dataset, we again see that given a full sibling relationship, cases have increased odds of going into the same main cluster (OR≈1.19-1.35, $p<0.00001$) and that clusters contain individuals with more similar genotype frequencies than the unclustered dataset (Average Fst≈0.13) (Table 2.8).



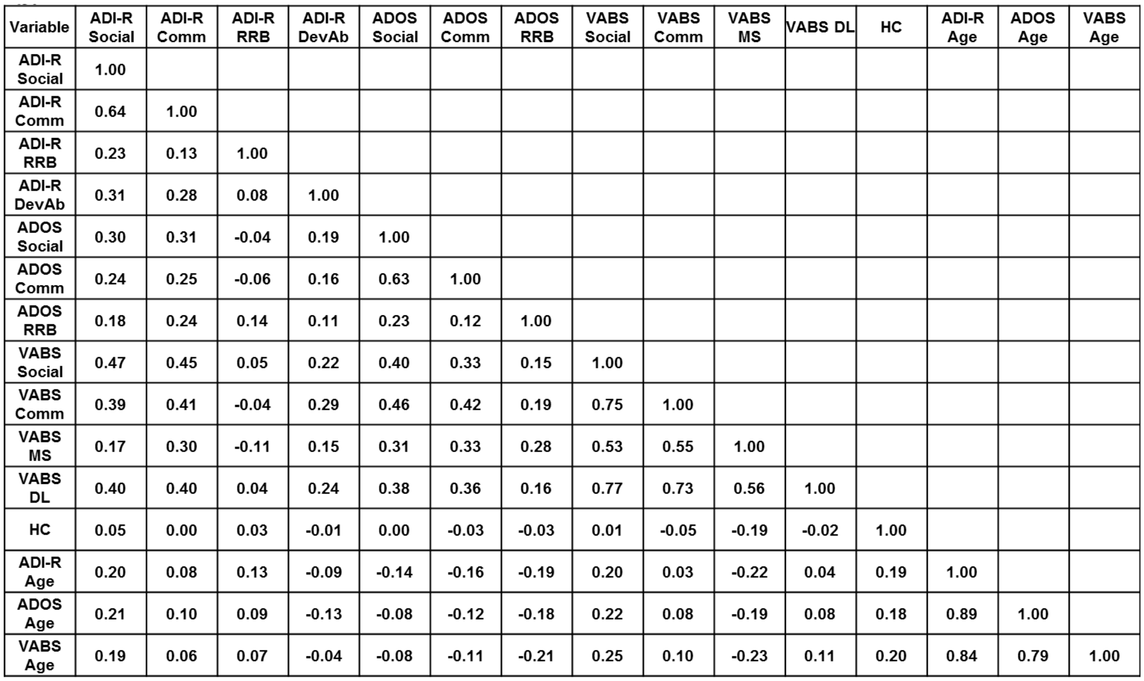
a.

Odds of Same Cluster Assignment Given Sibling Relationship		
Dataset	Odds Ratio Range	p-value
AGRE	1.38-1.42	<0.00001
AGP	1.19-1.35	<0.00001

b.

F-statistic Comparing Clusters to Unclustered Dataset			
Dataset	Mean Fst	Standard Error	95% Conf. Interval
AGRE	0.1664	$9.13x10^{-4}$	(0.1646, 0.1682)
$AGRE_{EA}$	0.1281	$7.97x10^{-4}$	(0.1265, 0.1296)
AGP	0.1251	$7.53x10^{-4}$	(0.1236, 0.1266)
AGP_{EA}	0.1031	$6.86x10^{-4}$	(0.1018, 0.1045)

Table 2.8. Results Evaluating Genetics Underlying Cluster Assignments. a. Odds Ratios represent increased odds of cases being assigned to the same cluster given a familial relationship. **b.** Average Wright's F-statistic (Fst) across informative autosomal markers comparing cluster subpopulations to total unclustered population. Fst reported for the entire clustering dataset and the European Americans (EA) only.



Variable	ADI-R Social	ADI-R Comm	ADI-R RRB	ADI-R DevAb	ADOS Social	ADOS Comm	ADOS RRB	VABS Social	VABS Comm	VABS MS	VABS DL	HC	ADI-R Age	ADOS Age	VABS Age
ADI-R Social	1.00														
ADI-R Comm	0.64	1.00													
ADI-R RRB	0.23	0.13	1.00												
ADI-R DevAb	0.31	0.28	0.08	1.00											
ADOS Social	0.30	0.31	-0.04	0.19	1.00										
ADOS Comm	0.24	0.25	-0.06	0.16	0.63	1.00									
ADOS RRB	0.18	0.24	0.14	0.11	0.23	0.12	1.00								
VABS Social	0.47	0.45	0.05	0.22	0.40	0.33	0.15	1.00							
VABS Comm	0.39	0.41	-0.04	0.29	0.46	0.42	0.19	0.75	1.00						
VABS MS	0.17	0.30	-0.11	0.15	0.31	0.33	0.28	0.53	0.55	1.00					
VABS DL	0.40	0.40	0.04	0.24	0.38	0.36	0.16	0.77	0.73	0.56	1.00				
HC	0.05	0.00	0.03	-0.01	0.00	-0.03	-0.03	0.01	-0.05	-0.19	-0.02	1.00			
ADI-R Age	0.20	0.08	0.13	-0.09	-0.14	-0.16	-0.19	0.20	0.03	-0.22	0.04	0.19	1.00		
ADOS Age	0.21	0.10	0.09	-0.13	-0.08	-0.12	-0.18	0.22	0.08	-0.19	0.08	0.18	0.89	1.00	
VABS Age	0.19	0.06	0.07	-0.04	-0.08	-0.11	-0.21	0.25	0.10	-0.23	0.11	0.20	0.84	0.79	1.00

Table 2.9. Variable Contributions to Principal Components of AGP Dataset. Variables contributing the most to the observed variance explained by each component are indicated in bold italics. Comm=Communication Domain Scores; RRB=Restricted, repetitive behaviors; DevAb=Abnormality of Development evident ≤36 months; MS=MotorSkills; DL=Daily Living; HC=head circumferences.

Phenotype Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10	Comp11	Comp12	Comp13	Comp14	Comp15
ADI-R Social	12.88	4.36	0.24	1.55	1.25	0.17	***19.75***	4.92	0.75	0.69	***52.69***	0.45	0.21	0.10	0.00
ADI-R Comm	13.00	0.96	0.40	0.61	1.90	0.46	***35.84***	2.68	1.03	1.10	***41.82***	0.17	0.00	0.03	0.00
ADI-R RRB	4.39	***21.55***	***50.27***	0.66	1.39	***19.17***	1.28	0.69	0.04	0.01	0.45	0.02	0.05	0.01	0.03
ADI-R DevAb	***21.19***	0.11	2.19	6.74	***57.40***	6.54	5.05	0.10	0.33	0.00	0.09	0.15	0.07	0.00	0.03
ADOS Social	10.08	1.53	4.22	0.52	3.64	10.89	7.94	8.05	1.62	***49.01***	2.07	0.06	0.13	0.14	0.10
ADOS Comm	7.25	1.92	5.43	0.15	2.66	***19.96***	7.68	8.07	4.74	***41.20***	0.73	0.12	0.04	0.06	0.00
ADOS RRB	***16.74***	6.26	17.82	***25.01***	5.29	***24.30***	3.70	0.18	0.00	0.57	0.12	0.00	0.00	0.00	0.00
VABS Social	4.53	0.55	1.89	0.57	2.99	0.00	0.11	16.33	3.97	0.07	0.00	0.01	15.36	***53.13***	0.49
VABS Comm	4.58	0.00	2.29	0.39	1.87	0.14	0.48	17.98	3.68	0.17	0.09	2.00	***66.24***	0.03	0.09
VABS MotorSkills	0.90	0.51	0.27	0.06	0.46	0.21	0.22	14.34	***72.78***	5.29	1.21	3.73	0.01	0.00	0.00
VABS DailyLiving	4.24	0.07	1.48	0.37	1.90	0.25	0.07	***23.12***	4.32	0.79	0.01	0.65	17.27	***44.94***	0.52
HC	0.00	***25.55***	11.00	***55.78***	6.56	0.53	0.26	0.30	0.00	0.01	0.01	0.00	0.01	0.00	0.00
ADI-R Age	0.12	14.13	0.45	2.73	3.99	6.53	4.56	1.52	2.95	0.02	0.05	8.59	0.10	0.27	***54.01***
ADOS Age	0.08	12.71	0.82	2.31	5.63	5.74	4.33	1.46	3.06	0.22	0.10	18.72	0.01	0.71	***44.11***
VABS Age	0.02	9.77	1.26	2.55	3.08	5.11	8.74	0.28	0.75	0.87	0.56	***65.33***	0.51	0.57	0.60

Table 2.10. Variable Contributions to Principal Components of AGP Dataset. Variables contributing the most to the observed variance explained by each component in the AGP dataset are indicated in bold italics.

Method	Measure	Dataset Partitions													
		2	3	4	5	6	7	8	9	10	11	12	13	14	15
agnes	Connectivity	***8.339***	249.971	270.538	368.010	368.010	400.844	480.350	525.765	628.859	695.811	695.811	811.612	835.692	917.338
	Dunn	0.149	0.155	0.155	0.141	0.141	0.141	0.149	0.149	0.149	0.156	0.156	0.171	0.171	***0.173***
	Silhouette	***0.297***	0.183	0.147	0.108	0.093	0.065	0.068	0.060	0.067	0.061	0.055	0.070	0.059	0.064
diana	Connectivity	385.861	713.150	717.686	927.514	1076.493	1154.714	1214.735	1346.593	1384.472	1422.792	1494.822	1511.073	1610.059	1676.895
	Dunn	0.107	0.112	0.116	0.118	0.120	0.120	0.122	0.127	0.131	0.134	0.134	0.134	0.124	0.126
	Silhouette	0.237	0.152	0.127	0.122	0.117	0.108	0.104	0.112	0.111	0.105	0.095	0.088	0.092	0.092
kmeans	Connectivity	413.851	632.644	801.270	843.503	1092.679	1067.820	1165.568	1181.950	1345.339	1461.656	1455.696	1418.996	1515.080	1569.366
	Dunn	0.068	0.124	0.111	0.078	0.090	0.090	0.111	0.085	0.132	0.103	0.107	0.137	0.105	0.107
	Silhouette	0.238	0.170	0.166	0.146	0.145	0.147	0.148	0.144	0.141	0.139	0.139	0.139	0.135	0.135
pam	Connectivity	407.137	864.928	814.154	1089.027	1179.318	1414.685	1371.974	1578.175	1671.510	1751.003	1657.883	1762.016	1890.952	1923.998
	Dunn	0.098	0.104	0.089	0.111	0.086	0.067	0.090	0.069	0.069	0.069	0.096	0.101	0.087	0.091
	Silhouette	0.235	0.144	0.162	0.152	0.142	0.131	0.127	0.117	0.114	0.112	0.119	0.115	0.114	0.112
sota	Connectivity	397.485	748.238	1033.950	1064.406	1199.766	1352.044	1487.189	1493.213	1546.187	1658.672	1774.110	1844.904	1922.660	1980.068
	Dunn	0.107	0.111	0.089	0.090	0.090	0.090	0.094	0.094	0.094	0.094	0.102	0.102	0.102	0.102
	Silhouette	0.237	0.150	0.121	0.119	0.106	0.109	0.107	0.102	0.096	0.093	0.095	0.097	0.097	0.096
clara	Connectivity	415.099	659.011	1062.616	1331.042	1409.810	1517.569	1627.525	1696.601	1813.571	1892.771	1944.889	2005.438	2074.149	2157.532
	Dunn	0.098	0.067	0.100	0.086	0.087	0.096	0.108	0.103	0.089	0.111	0.080	0.078	0.106	0.111
	Silhouette	0.196	0.151	0.114	0.109	0.102	0.117	0.113	0.107	0.100	0.098	0.099	0.107	0.095	0.096

Table 2.11. Validity Scores for Clustering: AGP Dataset. For each statistic, the best results observed for each cluster validation test are indicated in bold italics. These indicate the most valid partitions of the datasets, what method did this and number of separations needed. agnes=agglomerative hierarchical, diana=divisive hierarchical, pam=partitioning around medoids, sota=self-organizing tree algorithm, clara=clustering large applications. Dunn Index – smallest inter-cluster distance / largest intra-cluster distance, should be maximized (scale=0, ∞). Connectivity - to what extent are individuals placed in the same cluster as the most similar individuals, should be minimized (scale=0, ∞). Silhouette Width - overall average of average distance between individual and others in same cluster compared to different cluster (-1, 1); well-clustered=1.

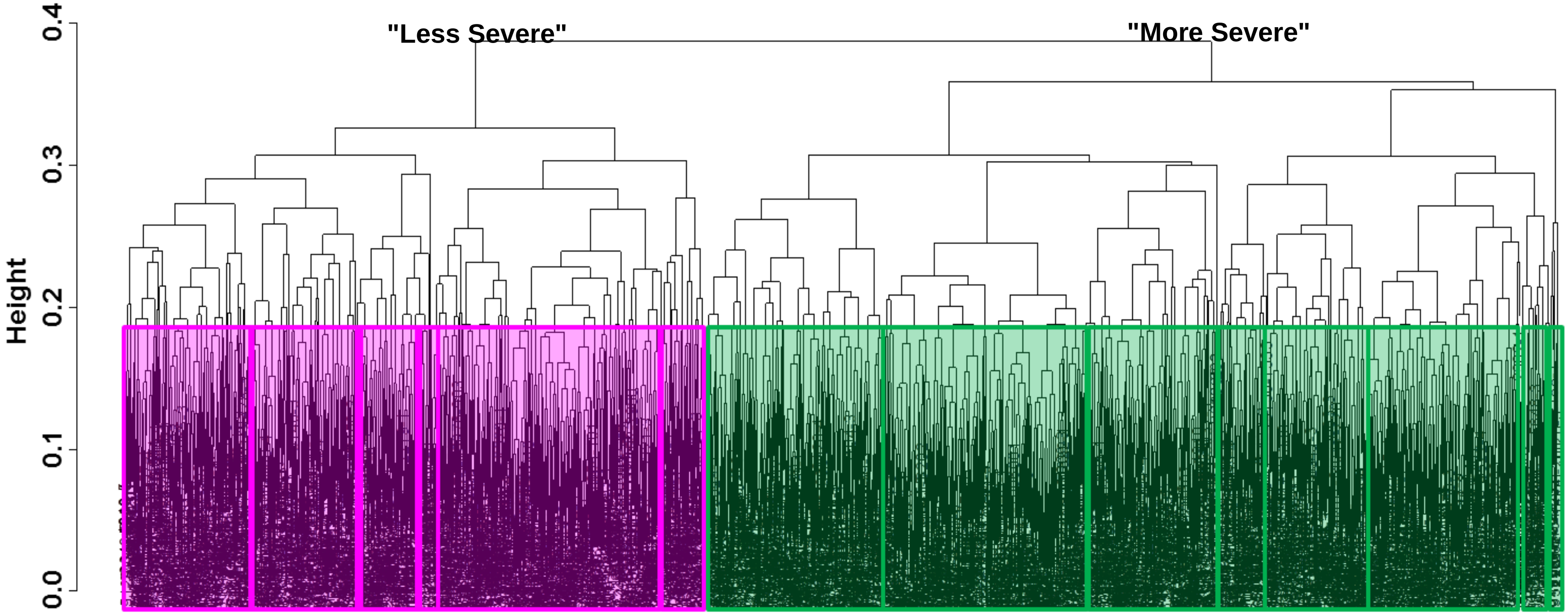
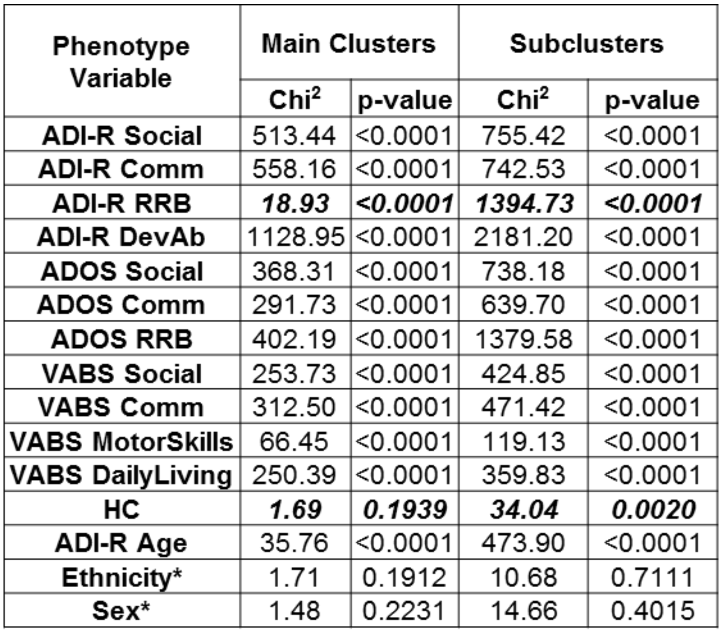
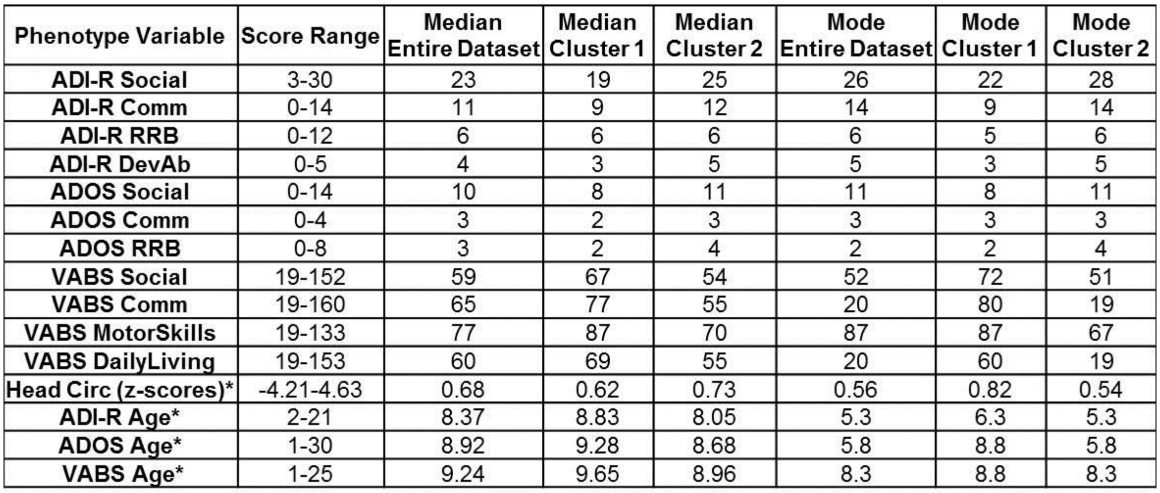


Figure 2.5. Agglomerative Clustering : AGP Dataset. Height indicates distance between merging clusters at successive stages of clustering and is related to dissimilarities among clusters. The main cluster highlighted in green represents individuals with more severe ASD phenotypes compared to individuals assigned to the main cluster highlighted pink. Subclusters are indicated with corresponding boxes. One subcluster only contained 2 individuals and is not denoted with a highlighted box.



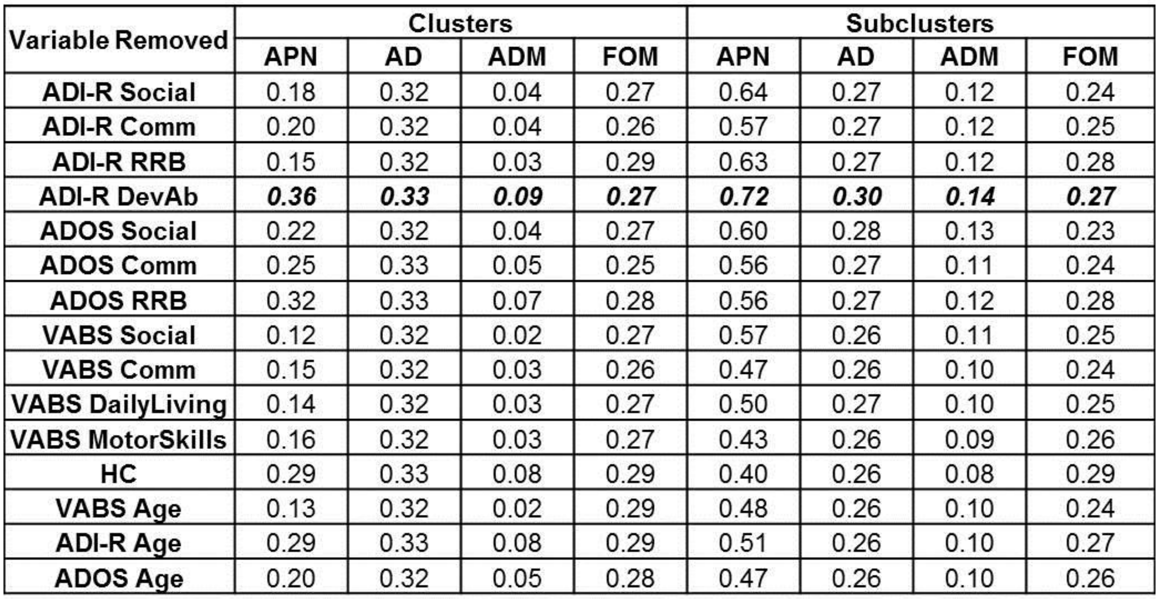
Phenotype Variable	Main Clusters		Subclusters	
	Chi^2	p-value	Chi^2	p-value
ADI-R Social	513.44	<0.0001	755.42	<0.0001
ADI-R Comm	558.16	<0.0001	742.53	<0.0001
ADI-R RRB	***18.93***	***<0.0001***	***1394.73***	***<0.0001***
ADI-R DevAb	1128.95	<0.0001	2181.20	<0.0001
ADOS Social	368.31	<0.0001	738.18	<0.0001
ADOS Comm	291.73	<0.0001	639.70	<0.0001
ADOS RRB	402.19	<0.0001	1379.58	<0.0001
VABS Social	253.73	<0.0001	424.85	<0.0001
VABS Comm	312.50	<0.0001	471.42	<0.0001
VABS MotorSkills	66.45	<0.0001	119.13	<0.0001
VABS DailyLiving	250.39	<0.0001	359.83	<0.0001
HC	***1.69***	***0.1939***	***34.04***	***0.0020***
ADI-R Age	35.76	<0.0001	473.90	<0.0001
Ethnicity*	1.71	0.1912	10.68	0.7111
Sex*	1.48	0.2231	14.66	0.4015

Table 2.12. Cluster Differences in AGP Dataset. Kruskal Wallis comparison of variable distributions between the two main clusters and across the 15 subclusters. All input variable distributions, except HC, are significantly different between the main clusters. HC distributions are significantly different across subclusters. Asterisks indicate information not used as input variable.



Phenotype Variable	Score Range	Median Entire Dataset	Median Cluster 1	Median Cluster 2	Mode Entire Dataset	Mode Cluster 1	Mode Cluster 2
ADI-R Social	3-30	23	19	25	26	22	28
ADI-R Comm	0-14	11	9	12	14	9	14
ADI-R RRB	0-12	6	6	6	6	5	6
ADI-R DevAb	0-5	4	3	5	5	3	5
ADOS Social	0-14	10	8	11	11	8	11
ADOS Comm	0-4	3	2	3	3	3	3
ADOS RRB	0-8	3	2	4	2	2	4
VABS Social	19-152	59	67	54	52	72	51
VABS Comm	19-160	65	77	55	20	80	19
VABS MotorSkills	19-133	77	87	70	87	87	67
VABS DailyLiving	19-153	60	69	55	20	60	19
Head Circ (z-scores)*	-4.21-4.63	0.68	0.62	0.73	0.56	0.82	0.54
ADI-R Age*	2-21	8.37	8.83	8.05	5.3	6.3	5.3
ADOS Age*	1-30	8.92	9.28	8.68	5.8	8.8	5.8
VABS Age*	1-25	9.24	9.65	8.96	8.3	8.8	8.3

Table 2.13. Summary Statistics for Unclustered vs Clustered AGP Datasets. Reported are medians and modes observed in the unclustered dataset compared to the two main clusters. Continuous variables are starred to indicate that the mean is reported in place of the median. Cases with scores indicating increased ASD severity preferentially cluster into the second, larger cluster. Age is reported in years.



Variable Removed	Clusters				Subclusters			
	APN	AD	ADM	FOM	APN	AD	ADM	FOM
ADI-R Social	0.18	0.32	0.04	0.27	0.64	0.27	0.12	0.24
ADI-R Comm	0.20	0.32	0.04	0.26	0.57	0.27	0.12	0.25
ADI-R RRB	0.15	0.32	0.03	0.29	0.63	0.27	0.12	0.28
ADI-R DevAb	***0.36***	***0.33***	***0.09***	***0.27***	***0.72***	***0.30***	***0.14***	***0.27***
ADOS Social	0.22	0.32	0.04	0.27	0.60	0.28	0.13	0.23
ADOS Comm	0.25	0.33	0.05	0.25	0.56	0.27	0.11	0.24
ADOS RRB	0.32	0.33	0.07	0.28	0.56	0.27	0.12	0.28
VABS Social	0.12	0.32	0.02	0.27	0.57	0.26	0.11	0.25
VABS Comm	0.15	0.32	0.03	0.26	0.47	0.26	0.10	0.24
VABS DailyLiving	0.14	0.32	0.03	0.27	0.50	0.27	0.10	0.25
VABS MotorSkills	0.16	0.32	0.03	0.27	0.43	0.26	0.09	0.26
HC	0.29	0.33	0.08	0.29	0.40	0.26	0.08	0.29
VABS Age	0.13	0.32	0.02	0.29	0.48	0.26	0.10	0.24
ADI-R Age	0.29	0.33	0.08	0.29	0.51	0.26	0.10	0.27
ADOS Age	0.20	0.32	0.05	0.28	0.47	0.26	0.10	0.26

Table 2.14. Sensitivity Analyses: AGP Dataset. Reported are results from sensitivity analyses. For the stability measures calculated, smaller values indicate more stable cluster results. Statistics evaluating cluster stability upon removal of each variable are: APN=Average proportion of nonoverlap or number of individuals not placed in same cluster when variable is removed (scale=0,1); AD= Average distance between individuals placed in same cluster when variable is removed (scale=0, ∞); ADM=Average distance between means between cluster centers for individuals placed in same cluster when variable is removed (scale=0, ∞); FOM=Figure of merit or average intra-cluster variance of the removed variable where clustering is based on remaining variables (scale=0, ∞).

Discussion

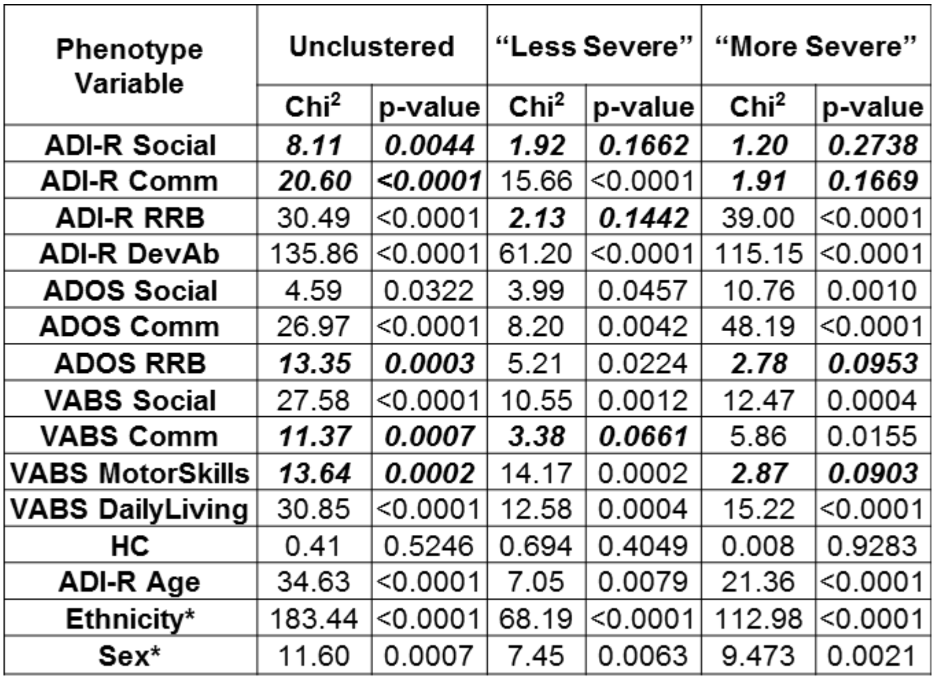
The extensive phenotypic variability within ASDs may hinder our ability to identify genotype-phenotype associations. To address this problem, we used multivariate statistical analyses to take advantage of ASD-related behavioral information from multiple sources and to include quantitative data relevant to macrocephaly. This approach allows effective evaluation of a broad array of data, enabling potentially more accurate phenotype definitions for large ASD datasets. We demonstrate that ASD phenotypic subgroups exist and can be replicated. Further, we demonstrate that these subgroups are genetically relevant.

Optimal Clustering Method

It is interesting that the optimal clustering method used to evaluate the ASD data is the agglomerative hierarchical method. This method uses connectivity based clustering and is unique from other methods because it begins with each individual as a separate cluster and aggregates them back together using the variable dissimilarities calculated for each individual when compared to every other individual[103]. It may be then for very complex traits, like those seen impaired in ASD, initially focusing on similarities across the dataset instead of differences will lead to identification of traits having the largest effect on variation overall.

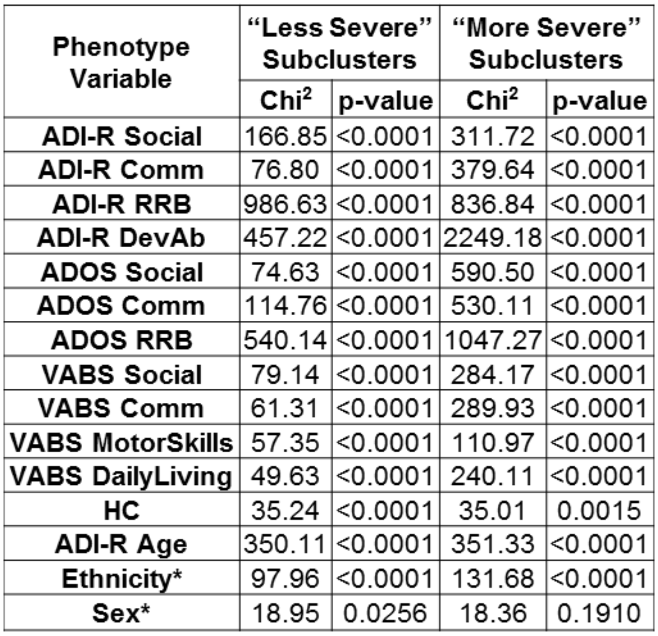
Phenotype Clusters

The strongest and most obvious clustering aggregates ASDs into two major clusters, grouped on overall symptom severity. When comparing variable distributions between the unclustered AGRE and AGP datasets, we observed that the two datasets had significantly different distributions of family structure and gender. The AGRE dataset having proportionally more females (z=3.41, p=0.003) and multiplex families (z=21.67, p<0.00001) than the AGP dataset. Previous research has suggested that phenotypic expression of ASD in multiplex families is distinct from that in simplex families[216]. There is also previous evidence indicating sex-specific effects in ASD[54]. It is striking that given these initial differences between datasets, our approach still identified main clusters with similar characteristics. In fact, some input variable distributions that were significantly different between the unclustered datasets were no longer significantly different when comparing distributions in the similarly-defined main clusters from both datasets (i.e. AGRE "less severe" and AGP "less severe") (Table 2.15). The initial differences between the AGRE and AGP datasets may account for the resulting subclusters being not as easily comparable (Table 2.16). Even with the different variable distributions we observe when comparing the subclusters from both datasets, there are some interesting similarities. For instance, at this level of subclustering we observe very small groups of cases that are remote from the other larger subclusters. In the AGRE dataset, we observe one small subcluster (n=47) within the 'less severe' main cluster, and three small subclusters (n=10, 38, 38) within the 'more severe' main cluster. The commonality across each of these smaller subclusters is that assigned individuals have large discrepancies between comparable domain scores (e.g., communication domain) from the ADI-R and ADOS. For example, individuals have either more severe scores on the ADI-R domains and less severe scores on the ADOS domains, when compared to larger subclusters within the same main cluster, or vice versa even though the ages at exam for both instruments are very similar. We see similar outlier groups in the AGP dataset subclusters.



Phenotype Variable	Unclustered		"Less Severe"		"More Severe"	
	Chi^2	p-value	Chi^2	p-value	Chi^2	p-value
ADI-R Social	***8.11***	***0.0044***	***1.92***	***0.1662***	***1.20***	***0.2738***
ADI-R Comm	***20.60***	***<0.0001***	15.66	<0.0001	***1.91***	***0.1669***
ADI-R RRB	30.49	<0.0001	***2.13***	***0.1442***	39.00	<0.0001
ADI-R DevAb	135.86	<0.0001	61.20	<0.0001	115.15	<0.0001
ADOS Social	4.59	0.0322	3.99	0.0457	10.76	0.0010
ADOS Comm	26.97	<0.0001	8.20	0.0042	48.19	<0.0001
ADOS RRB	***13.35***	***0.0003***	5.21	0.0224	***2.78***	***0.0953***
VABS Social	27.58	<0.0001	10.55	0.0012	12.47	0.0004
VABS Comm	***11.37***	***0.0007***	***3.38***	***0.0661***	5.86	0.0155
VABS MotorSkills	***13.64***	***0.0002***	14.17	0.0002	***2.87***	***0.0903***
VABS DailyLiving	30.85	<0.0001	12.58	0.0004	15.22	<0.0001
HC	0.41	0.5246	0.694	0.4049	0.008	0.9283
ADI-R Age	34.63	<0.0001	7.05	0.0079	21.36	<0.0001
Ethnicity*	183.44	<0.0001	68.19	<0.0001	112.98	<0.0001
Sex*	11.60	0.0007	7.45	0.0063	9.473	0.0021

Table 2.15. Dataset Differences. Kruskal Wallis comparisons of variable distributions between the AGRE and AGP datasets, as well as the resulting clusters. Particularly, ADI-R social scores, and ADI-R & ADOS RRB scores are more divergent between the two datasets than the comparable main clusters. Unclustered=AGRE vs. AGP dataset; "Less Severe"="less severe" AGRE cluster vs. "less severe" AGP cluster and similar for the "more severe" clusters. Asterisks indicate information not used as input variable.



Phenotype Variable	"Less Severe" Subclusters		"More Severe" Subclusters	
	Chi^2	p-value	Chi^2	p-value
ADI-R Social	166.85	<0.0001	311.72	<0.0001
ADI-R Comm	76.80	<0.0001	379.64	<0.0001
ADI-R RRB	986.63	<0.0001	836.84	<0.0001
ADI-R DevAb	457.22	<0.0001	2249.18	<0.0001
ADOS Social	74.63	<0.0001	590.50	<0.0001
ADOS Comm	114.76	<0.0001	530.11	<0.0001
ADOS RRB	540.14	<0.0001	1047.27	<0.0001
VABS Social	79.14	<0.0001	284.17	<0.0001
VABS Comm	61.31	<0.0001	289.93	<0.0001
VABS MotorSkills	57.35	<0.0001	110.97	<0.0001
VABS DailyLiving	49.63	<0.0001	240.11	<0.0001
HC	35.24	<0.0001	35.01	0.0015
ADI-R Age	350.11	<0.0001	351.33	<0.0001
Ethnicity*	97.96	<0.0001	131.68	<0.0001
Sex*	18.95	0.0256	18.36	0.1910

Table 2.16. Subcluster Differences between Datasets. Kruskal Wallis comparisons of variable distributions across subclusters from both datasets. All input variable distributions are significantly different among the subclusters when comparing these groups between the AGRE and AGP datasets. "Less Severe" = Subclusters within the "less severe" main clusters compared between datasets and similar for the "more severe" subclusters. Asterisks indicate information not used as input variable.

With the exception of these small outlier subclusters, it is apparent that age ranges are much smaller within subclusters when compared to main clusters. In both the AGRE and AGP datasets, there is one subcluster grouped separately from the other subclusters within the 'more severe' main cluster that contains some of the youngest individuals in the datasets ($\bar{x}_{AGRE\_ADI\text{-}R}$=6.6 years, 95%$CI_{AGRE\_ADI\text{-}R}$=6.3-6.9, $n_{AGRE\_ADI\text{-}R}$=416; $\bar{x}_{AGP\_ADI\text{-}R}$=5.7 years, 95%$CI_{AGP\_ADI\text{-}R}$=5.4-6.1, $n_{AGP\_ADI\text{-}R}$=277). In the AGRE dataset there are also two subclusters within the 'more severe' main cluster that include a majority of nonverbal individuals (61%-63% nonverbal) when compared to other subclusters within the 'more severe' main cluster (0-18% nonverbal) and the subclusters comprising the 'less severe' main cluster (6%-14% nonverbal). We also see two similar subclusters within the 'more severe' AGP dataset cluster (61%-64% nonverbal).

We see that scores assessing similar ASD traits do not correlate strongly between the ADI-R and ADOS, especially with regard to the RRB measure, even though all individuals evaluated meet ASD diagnostic criteria on both instruments. Our observations of weaker correlations for the RRB measures are also consistent with other studies where weaker correlation was observed between the ADI-R and ADOS repetitive behavior scores compared to the social and communication scores[118]. Previous studies have also shown that the ADI-R and ADOS make independent, additive contributions to more accurate diagnostic decisions and that specificities improve significantly when both instruments are used compared to each alone[105]. Our results provide further evidence that including information from both tests is important for precise definition of ASD phenotypes.

Effect of Developmental Abnormality Measure

All variables included as input in our multivariate analyses influence PCA results and cluster assignment. However, the 'severity of abnormalities related to ASD behavioral criteria exhibited by 36 months of age' (DevAb) score from the ADI-R stands out as having a stronger influence on cluster and PCA results. This measure is used in diagnosis in the ADI-R, based on criteria established by the DSM-IV. There must be evidence of deficient social or communication skills prior to or by 36 months for a diagnosis of strict autism to be made[128]. We see that this measure from the ADI-R does not exhibit strong correlations with any other input variable and has a substantial influence on the phenotypic variance explained in the first PC of both datasets. This measure has consistently different distributions between clusters and across subclusters and the largest overall effect on cluster and subcluster stability. In both the discovery and replication datasets, we see in the resulting 'more severe' clusters that 59-80% of individuals received the highest score possible for this measure indicating very severe abnormality of development observed early in life, compared to 0-0.4% of individuals in the 'less severe' clusters.

Effect of Repetitive Behavior Measures

Repetitive behaviors also stand out from other variables in their contribution to the phenotypic variance explained in the first three PCs of both datasets. ADOS RRB measures have a strong contribution to the first data component, and consequently have significantly different score distributions between individuals in the two main clusters. Interestingly, ADI-R RRB measures are not strong contributors to the first PC of the AGRE dataset. However, ADI-R RRB measures do have strong contributions to PC2 and PC3. In the AGP dataset, the contribution from these measures to PC1 is more comparable to other input variables. Yet, ADI-R RRB scores still do not contribute as much to PC1 as do RRBs assessed with the ADOS. This is also apparent in the clustering results; ADI-R RRBs are not significantly different between the two main AGRE dataset clusters but are significantly different between the two main AGP dataset clusters. These scores also have significantly different distributions across the subclusters from both datasets. It is interesting that RRB measures have different levels of influence on both phenotypic variance defined via PCA and definition of the two main clusters, based on whether they are evaluated with the ADOS or the ADI-R.

One explanation for the differing influence of RRBs on multivariate statistical results when comparing diagnostic instruments is that RRBs are not as extensively evaluated with the ADOS as with the ADI-R. RRBs observed on the ADOS are more likely to be simple repetitive behaviors that are easily observed in a brief interaction. Many RRBs are difficult to assess in a short period of time because certain restrictive and repetitive behaviors may only occur in specific situations, and the ADOS is limited by both time and context[95]. In contrast, the ADI-R captures a broader array of RRBs and provides information for more complex repetitive behaviors. It is notable that by including ADI-R domain scores and not item level scores we are not fully distinguishing simple versus complex repetitive behaviors.

An explanation for the differing influence of ADI-R RRBs on multivariate statistical results when comparing datasets is that the AGP dataset has more than twice the number of individuals with this information than does the AGRE dataset. Since the ADI-R is useful for distinguishing types of RRBs, it may be necessary to have more data from individuals exhibiting similar RRB characteristics for this measure to have an appreciable impact on main cluster assignment. Even with this difference, ADI-R RRB scores are more noticeably distinct across the subclusters when compared to the main clusters from both datasets.

The combined evidence from PCA and agglomerative hierarchical clustering suggest that presence of RRBs is important to ASD phenotype definitions in these datasets and that this behavior is unique from the social and communication deficits for definition of ASD subphenotypes. This is in line with numerous previous studies[29, 36, 82, 95, 138, 163, 172, 173, 182, 188]. There is also evidence that ADI-R RRB scores have the strongest within-family concordance when compared to the social and communication measures providing support for a uniquely inherited component[197].

Effect of Head Circumference

Head circumferences do not contribute significantly to the phenotypic variance observed in the first principal component of either dataset, which by design defines more variance than any other PC[87]. We also see that the distributions for this measure are not significantly different between the two main clusters grouped by overall ASD severity. We do, however, see a substantial contribution to the definition of phenotypic variance explained by the third PC of the AGRE dataset and the second PC of the AGP dataset, and HCs do seem to have a strong influence on subcluster assignment. However, in the AGRE dataset HCs are significantly different across the subclusters regardless of main cluster assignment whereas in the AGP dataset, HCs are only significantly different across the subclusters comprising the less severe main cluster. We were surprised that head circumference did not have a stronger influence on main cluster assignment, AGP subcluster assignment, and definition of PC1. It is notable that for both evaluated datasets, the mean normalized HC is above average compared to individuals not diagnosed with a spectrum disorder ($\bar{x}_{\mathrm{AGRE}}$=0.72, $\bar{x}_{\mathrm{AGP}}$=0.66). It is possible that most individuals with ASD have larger head circumferences compared to normal individuals and that this is not a distinguishing trait for ASD subgroups but rather a trait specific to the broader diagnostic classification. Macrocephaly roughly defined as >2 standard deviations above the mean is only comorbid in ~13% of individuals for which this measure is available, in both the AGRE and AGP datasets. These rates are slightly lower than expected based on previously reported estimates ranging from 15-35%[62, 227]. This is consistent with other observations indicating individuals with ASD have increased head growth but do not meet criteria for macrocephaly. Unfortunately, HC measures are only available for ~54% of the AGRE dataset and ~47% of the AGP dataset. This could also be an explanation for the observed impact of HC on cluster assignment and the lower rate of macrocephaly in the AGRE and AGP datasets.

Another important caveat to our evaluation of head circumference is that ethnicity is noted to be important in head circumference normalization[227]. We normalized HC measures using a non-diseased population of European descent, due to our inability to identify normal population statistics for other ethnicities of interest with a similar age range to the datasets evaluated in our study. Our datasets have a slightly different ethnic background than does the normal population we used to normalize HC and this could affect our z-score calculations. We also did not take into account height, another factor that should be considered when evaluating macrocephaly, since this information is available for even fewer individuals in the AGRE and AGP datasets (~29% and ~46%, respectively). Height and head circumference measures, when available in our datasets, do exhibit positive correlations suggesting the increased HC may be due to increased stature and not necessarily exhibition of an endophenotype (ρ_{AGRE}=0.66; ρ_{AGP}=0.44) (Figure 2.6).

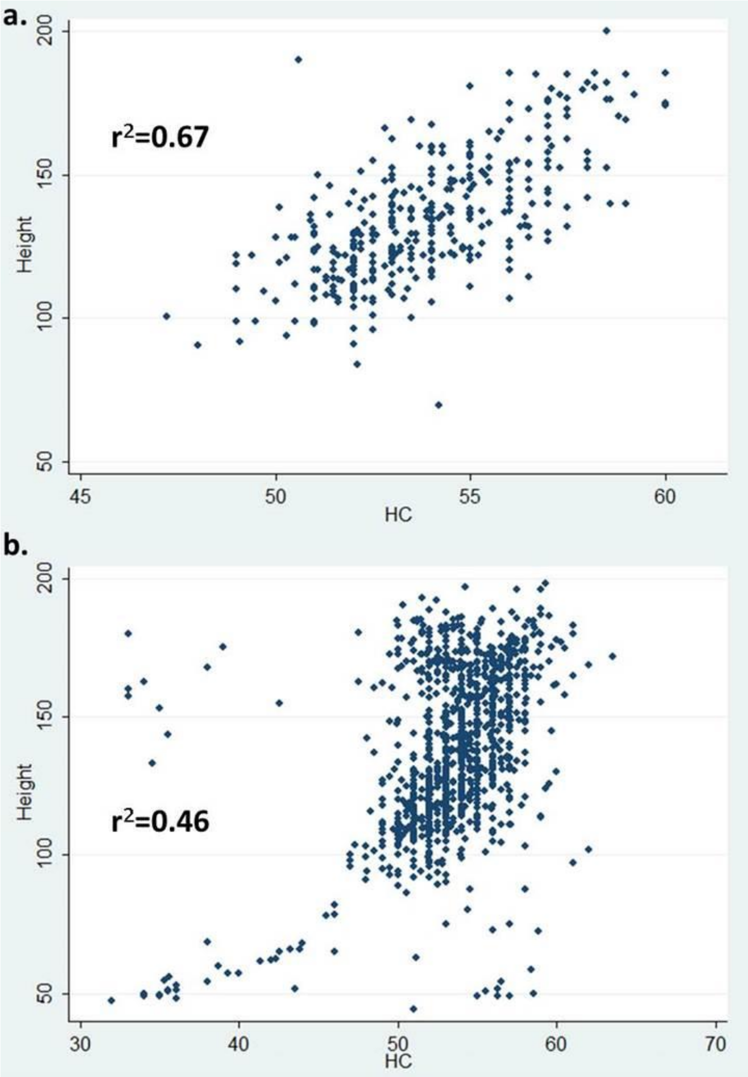


Figure 2.6. Correlation of Head Circumference & Height in Evaluated Datasets. Plotted are head circumferences (cm) versus height (cm) for individuals in the **a.** AGRE dataset and **b.** AGP dataset. Reported are squared Pearson's correlation coefficients (r^2).

The combined results from PCA and clustering indicate HC is important in defining ASD subphenotypes, but not in determining overall severity. Again, these measures are not available for a large portion of the cases in the datasets we evaluated, which could affect the variable's impact on definition of PC1 and main clusters even with our stringent weighting scheme and the ability of the methods to allow for missing data. The same is also true for Vineland domain standard scores. While these scores do seem to be involved in definition of the main clusters, they do not contribute greatly to PC definition or stand out as classifying variables. These findings are possibly a result of having fewer individuals with VABS scores compared to ADI-R and ADOS scores. We chose to retain cases that are missing VABS and HC information since these are not considered ASD-specific diagnostic criteria.

Familial Clustering

Odds ratios showed significantly increased odds for affected siblings to cluster together into the two main clusters when compared to unrelated cases. These calculations are indicative of underlying genetic architecture. Further supporting this assumption, Wright's Fst calculations suggest cases with more similar genetic architecture clustered together into the two main clusters. Although Fst can be confounded by genetic ancestry, we obtained similar results using only individuals with European ancestry. It is notable that there is still evidence for significant genetic and phenotypic heterogeneity within ASD families. This is in agreement with many previous studies and the growing body of evidence reporting the involvement of *de novo* mutations arising in the germ-line[25, 176, 177]. However, the relationship of genotype to phenotype should be somewhat independent of inheritance patterns. While our results supported an underlying genetic influence on overall cluster assignment, to determine the true contribution of genetic factors to phenotypic cluster assignment it will be necessary to perform future genetic analyses based on these cluster groupings.

The overlapping interpretation of our results from two different multivariate analyses, PCA and clustering, demonstrate the utility of this approach. That we were able to show defined subgroups of phenotypic expression appearing to be genetically meaningful in the AGRE dataset and replicate these findings in an independent AGP dataset lends further support to the validity of the resulting cluster groupings and the idea that the phenotype clusters recapitulate underlying genetic mechanisms in Autism Spectrum Disorders.

CHAPTER III

PATHWAY-BASED GENOME-WIDE ASSOCIATION STUDIES IN DEFINED SUBGROUPS

Introduction

Genetic factors have a strong influence in the etiology of ASD. However, the individual effects of most previously implicated common variants are modest, tend not to replicate in independent cohorts, and the combined evidence from many analyses does not explain the estimated heritability[21, 44, 45, 79, 159]. The difficulty in identifying common, inherited variation with replicable effects may arise from the wide variability in clinical manifestation of ASD and the relationship to genetic influences.

Evaluating larger sample sizes is one way to increase power in genetic studies of complex disorders, like ASD[56]. Studies have been conducted in large ASD cohorts when the phenotype is categorized dichotomously (i.e., affected/unaffected) by diagnostic cut-offs[17, 123, 230]. However, none of these associations replicate in independent cohorts, suggesting an increase in sample size is not sufficient to optimize power for ASD. It is also difficult to interpret potential phenotype-genotype relationships using results from these large-scale genetic analyses since evaluated cases express a wide continuum of symptom severity.

Previous studies have defined more phenotypically homogenous subgroups in ASD using overall trait severity, endophenotypes, and comorbidity information and evaluated genetic contributions to these subgroups[7, 24, 88, 90, 179, 183, 184, 204]. In numerous cases, linkage and association signals were increased despite a substantial reduction in sample size. Many of these studies also replicate previous results from analyses performed when subphenotypes of ASD were further defined[8, 40, 130]. These studies provide strong support for phenotypic subgroups being genetically meaningful.

Factors further complicating genetic association studies in ASD are related to the complexity of the underlying genetic models of the disorder. There are hundreds of different genes and risk loci implicated in ASD etiology[26, 143, 157]. Few, if any, of the currently identified genetic factors alone seem to contribute strong effects (OR>1.2) to risk for ASD[16, 50]. These smaller effects are easily overlooked in the typical approach to analysis of GWAS data, which looks for the most significantly associated individual single nucleotide polymorphisms (SNPs). Evidence from multiple independent studies indicates common, inherited variants have a cumulative effect on ASD risk[68, 107, 168]. In reality, genes often work as complex interacting networks, especially those involved in neural development. Pathway-based analysis of genome-wide SNP data considers the combined effects of multiple genetic variants functioning together in biological pathways[155, 218]. By applying this methodology to analysis of ASD genetic data, causal pathways and/or genetic interactions may be implicated giving biological insights that would otherwise be imperceptible[18, 86, 121, 132, 158].

Our hypothesis is that performing pathway-based genetic analyses in more phenotypically homogeneous ASD subgroups accounts for some heterogeneity, thus increasing power to detect genetic effects. We previously performed extensive phenotypic analyses in an Autism Genetic Resource Exchange (AGRE)[72] dataset[212]. We used data from the Autism Diagnostic Interview-Revised (ADI-R)[128], Autism Diagnostic Observation Schedule (ADOS)[127], Vineland Adaptive Behavior Scales[196], head circumferences, and ages as classifying variables. Unsupervised clustering identified two distinct groups of cases, dividing primarily on the severity of phenotypes. The same approach similarly identified two distinct groups of cases and confirmed this severity-based dichotomy in an independent dataset from the Autism Genome Project (AGP)[93].

In addition, there was significant familial clustering within groups (OR≈1.38-1.42, $p<0.00001$), suggesting that the clusters recapitulated genetic etiology. Identifying biological pathways and sets of genes contributing to the underlying mechanisms involved in expression of subphenotypes of ASD will help us gain further insight into the functional foundations of the various phenotypic aspects of this disorder. This study is one of the first to apply pathway analysis to ASD GWAS data, and to apply this methodology to well-defined subgroups of affected individuals.

Methods

Dataset Demographics and Quality Control

The discovery dataset consisted of individuals from the *AGRE* family-based study. We used previously generated, publicly available genetic data; samples were genotyped on the Illumina Bead Array and Affymetrix 550 chip[134]. Genetic data were merged in PLINK[166] and the final merged datasets were subjected to numerous quality control (QC) procedures (Figure 3.1). The final discovery dataset included 4,110 individuals (2,559 males and 1,551 females) in 895 families. 91.2% of these families were multiplex, 8.6% were simplex, and 0.2% had unknown family structure. Genetic ancestry determined by the software program Structure[165] was 80.1% European American, 16.2% Mexican American, 2.8% African American, and 0.8% mixed ancestry. After QC, a total of 507,669 SNPs, with a genotyping rate of 99.4%, were analyzed in discovery association analyses.

For the validation dataset, we used samples from the AGP database. Samples were previously genotyped on the Illumina 1M platform[134]. The same QC procedures used on genotyping data from the discovery dataset were used for the validation dataset. The final validation dataset contained 8,908 individuals (5,475 males and 3,275 females) in

2,960 families. 31% of the families in the validation dataset were multiplex, 49% were simplex families, and 20% of the dataset had unknown family structure. Genetic ancestry was 91.4% European American, 5.8% Mexican American, 2.6% African American, 0.2% mixed ancestry. After QC, a total of 779,343 SNPs with a genotyping rate of 99.8%, were analyzed in validation association analyses.

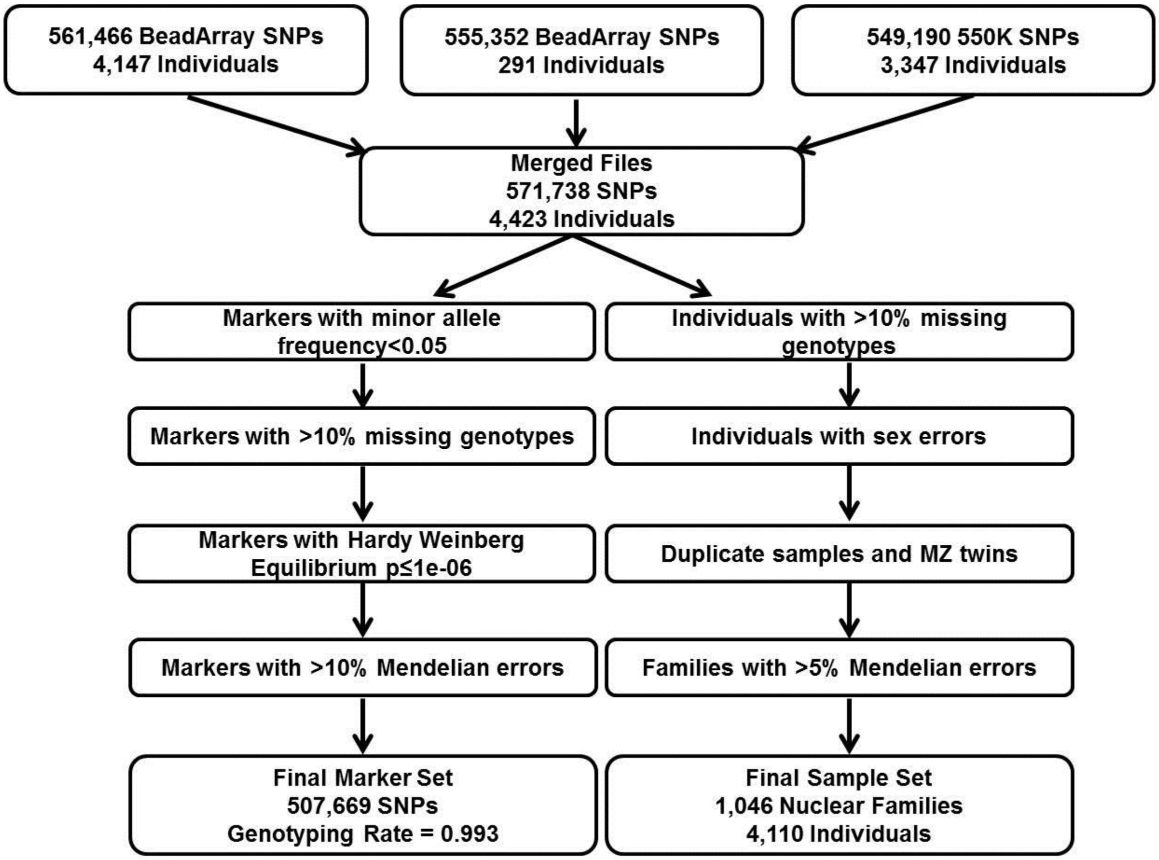

561,466 BeadArray SNPs
4,147 Individuals
555,352 BeadArray SNPs
291 Individuals
549,190 550K SNPs
3,347 Individuals
Merged Files
571,738 SNPs
4,423 Individuals
Markers with minor allele frequency<0.05
Individuals with >10% missing genotypes
Markers with >10% missing genotypes
Individuals with sex errors
Markers with Hardy Weinberg Equilibrium p≤1e-06
Duplicate samples and MZ twins
Markers with >10% Mendelian errors
Families with >5% Mendelian errors
Final Marker Set
507,669 SNPs
Genotyping Rate = 0.993
Final Sample Set
1,046 Nuclear Families
4,110 Individuals

Figure 3.1. Quality Control Procedures. Outlined is a flow diagram detailing exclusion criteria used to obtain quality genotyping data for AGRE discovery analyses and the final number of evaluated SNPs and samples. Marker exclusion criteria are detailed on the left and sample exclusion criteria are detailed on the right.

Single-SNP Association Analyses: AGRE Dataset

We used the AGRE family dataset for our initial modeling. Exclusion criteria and affection status for association analyses were selected based upon phenotype analyses described in detail in Chapter II. Briefly, cases meet Diagnostic and Statistical Manual-IV (DSM-IV) criteria for an Autism Spectrum Disorder diagnosis on both the Autism

Diagnostic Interview-Revised[128] (ADI-R) and the Autism Diagnostic Observation Schedule[127] (ADOS), age at ADI-R 2-21 years old. We excluded individuals with potential non-idiopathic autism (e.g. known neurogenetic disorders, known chromosomal abnormalities, prematurity <35 weeks). We used agglomerative hierarchical clustering to group individuals with ASD relative to multiple sources of behavioral and clinical exam information. Association analyses were performed using the Family-Based Association Test (FBAT)[114]. We tested the null hypothesis of no association in the presence of linkage using the empirical variance-covariance estimator under an additive, multi-allelic genetic model[113]. We performed three FBAT analyses for the AGRE dataset according to the phenotypic subgrouping (Figure 3.2).

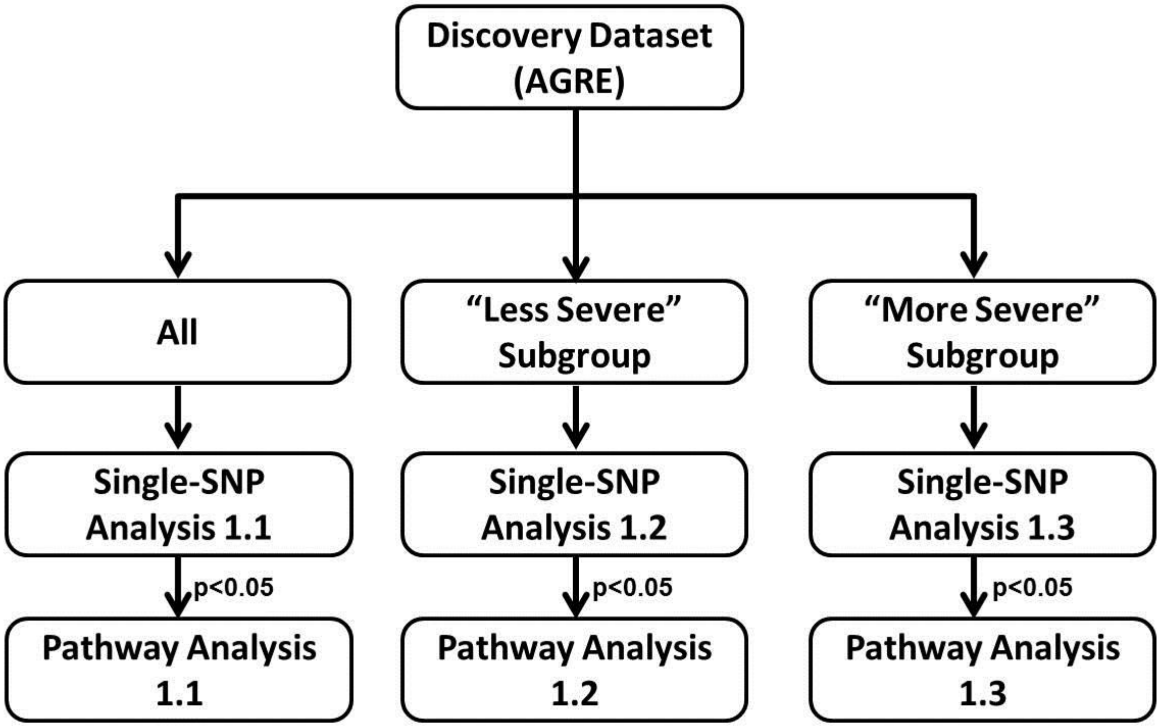
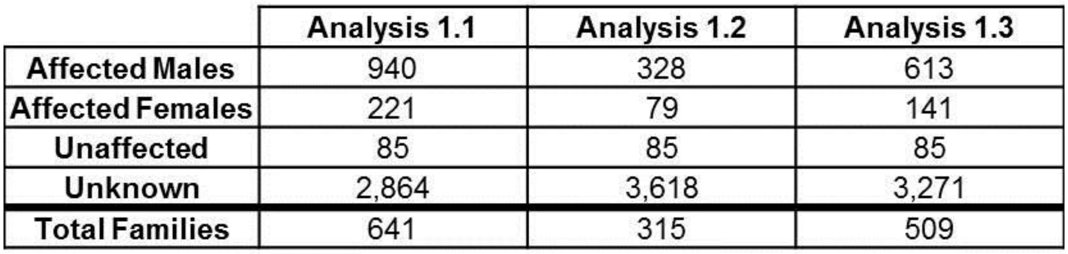


Figure 3.2. Analysis Plan Schematic: AGRE Dataset. Three single-SNP association analyses and subsequent pathway analyses were performed on the discovery dataset based on different ASD phenotype definitions. All=no phenotypic subgrouping; 'Less Severe'=individuals in the less severe subgroup; 'More Severe'=individuals in the more severe subgroup.

For the first analysis, affection status was assigned to all individuals meeting criteria for an ASD diagnosis on both the ADI-R and the ADOS, regardless of phenotypic subgrouping. There were 48 males and 37 females in this dataset that were evaluated on both the ADI-R and ADOS and did not meet diagnostic criteria for an ASD on either instrument. These individuals were coded as unaffected. We also analyzed the data with the 85 unaffected individuals alternatively coded as unknown and compared FBAT results at each SNP. There were no differences in p-value for evaluated SNPs. For the second analysis, affection status was assigned only to individuals in a ‘less severe’ subgroup. Cases in this subgroup have scores indicating less severe measures for interrogated behavioral and clinical exam information. Cases assigned to the alternate subgroup were coded with unknown affection status. For the third analysis, affection status was assigned only to individuals in a ‘more severe’ subgroup. Cases in this subgroup have scores indicating more severe measures for interrogated behavioral and clinical exam information (Table 3.1).



	Analysis 1.1	Analysis 1.2	Analysis 1.3
Affected Males	940	328	613
Affected Females	221	79	141
Unaffected	85	85	85
Unknown	2,864	3,618	3,271
Total Families	641	315	509

Table 3.1. Breakdown of Affection Status for Single-SNP Analyses: AGRE Dataset. Reported are the numbers of individuals evaluated for informative transmissions in Family-based Association Tests.

Deviation from the expected chi-square distribution was visualized in quantile-quantile plots generated with a unique source code and the ggplot2 package in R[210, 224]. Population substructure does not cause type I error in family-based association tests, however, due to the diverse genetic ancestry of the evaluated dataset, genomic inflation factors (λ) were estimated for results from each FBAT analysis using the GenABEL package for R[20]. Manhattan plots were produced using a unique source code and the ggplot2 package in R[210, 224]. The estimation of odds ratios (ORs) and 95% confidence interval (CI) calculations for evaluated SNPs were performed using UNPHASED[52]. To determine the overall OR for genes of interest, an average was calculated for all SNPs located within each gene boundary.

Pathway Analyses: AGRE dataset

Three separate pathway analyses were performed with the Pathway Analysis by Randomization Incorporating Structure (PARIS) pathway analysis software package[229] using p-values generated in the corresponding single-SNP analysis (i.e. 'Analysis 1.1', '1.2', '1.3'). By assigning SNPs to genes based on chromosomal locations and looking for functionally-defined gene sets with an overrepresentation of significant SNPs, PARIS identifies biological pathways of interest. Since we expected that there would be many variants of minor effect working together, we set a less-stringent threshold ($p<0.05$) for SNPs entered into the subsequent pathway analyses to ensure this information was captured. We evaluated 209 pathways defined in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for pathway-based association[101]. SNPs were assigned to a pathway gene if it fell within +/−50 kb of the ENSEMBL genomic interval (build hg19). Hapmap CEPH samples (release 27) were used to account for patterns of linkage disequilibrium (LD). Bonferroni corrected significance for evaluated pathways was $p\leq0.0002$. However, PARIS was currently only designed to generate pathway p-values as low as $p<0.001$. Also, many KEGG pathways contained overlapping genes and each significance test was not independent. Therefore, we chose the most stringent significance threshold available ($p<0.001$) for pathway-based results.

Since the primary functional focus of pathways defined in the KEGG database is not neurodevelopment, in order to more thoroughly understand the relationship of identified pathways to ASD we felt it was necessary to further subject significantly associated

KEGG-defined pathways to permutation-based investigations and identify ‘core’ genes driving overall pathway associations. We defined ‘core’ genes as genes whose p-value in the context of the biological pathway was p<0.001, and upon removal from pathway analysis reduced the significance of the overall pathway above the significance threshold. To determine the overall OR for 'core' genes, an average was calculated for all SNPs located within the gene boundary, while taking into account the direction of the effect

Single-SNP Association and Pathway Analyses: AGP Dataset

We performed three FBAT validation analyses in the AGP dataset similar to that described above for the AGRE dataset (Figure 3.3). The breakdown of affection status for subgroup-specific single-SNP analyses is reported in Table 3.2.

Nominally significant SNPs (p<0.05) from each of the single-SNP association analyses were subsequently evaluated in respective pathway analyses, via PARIS, as described above for the AGRE dataset (Figure 3.3).

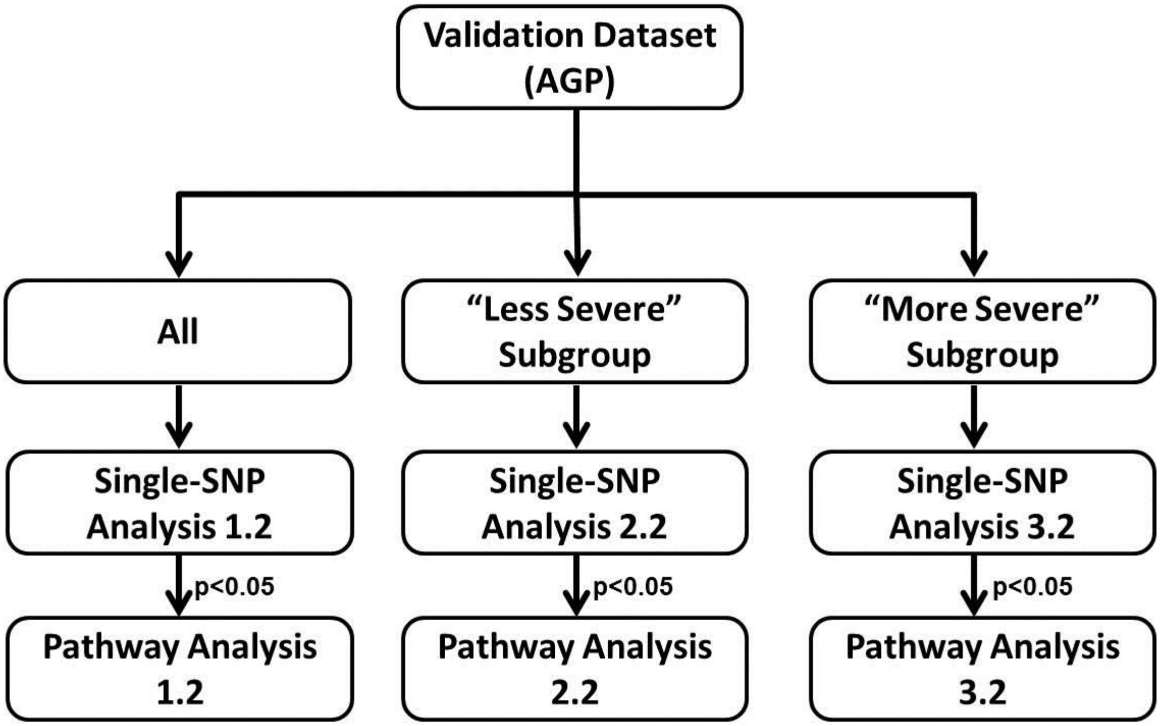
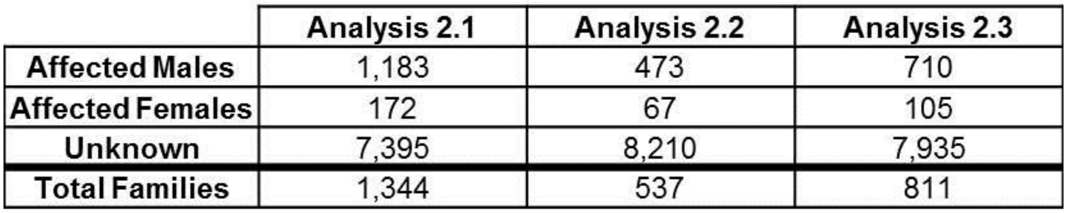


Figure 3.3. Analysis Plan Schematic: AGP Dataset. Three single-SNP association analyses and subsequent pathway analyses were performed on the validation dataset based on different ASD phenotype definitions. All=no phenotypic subgrouping; 'Less Severe'=individuals in the less severe subgroup; 'More Severe'=individuals in the more severe subgroup.



	Analysis 2.1	Analysis 2.2	Analysis 2.3
Affected Males	1,183	473	710
Affected Females	172	67	105
Unknown	7,395	8,210	7,935
Total Families	1,344	537	811

Table 3.2. Breakdown of Affection Status for Single-SNP Analyses: AGP Dataset. Reported are the numbers of individuals evaluated for informative transmissions in Family-based Association Tests using the AGP dataset.

Results

Single-SNP Association Analyses: AGRE Dataset

A total of 507,675 SNPs were analyzed for association in the discovery analyses. These SNPs were evaluated for association with all individuals meeting diagnostic criteria for an ASD on both the ADI-R and ADOS (Single-SNP Analysis 1.1), only affected individuals assigned to the 'less severe' phenotypic subgroup (Single-SNP Analysis 1.2), and only affected individuals assigned to the 'more severe' ASD subgroup (Single-SNP Analysis 1.3) (Figure 3.2). Genomic inflation factors for these analyses were 1.028, 1.020, and 1.011, respectively (Figure 3.4). This indicates that population structure had no appreciable impact on our results[49]. No SNPs met a Bonferroni corrected significance threshold of $p \leq 9.85 \times 10^{-8}$ for any of the three association analyses (Figure 3.5).

From Single-SNP Analysis 1.1, there were 26,970 SNPs ($p<0.05$) further evaluated in Pathway Analysis 1.1. From Single-SNP Analysis 1.2, 26,712 SNPs were evaluated in Pathway Analysis 1.2 and from Single-SNP Analysis 1.3, 26,335 SNPs were evaluated in the Pathway Analysis 1.3. Only 655 SNPs were associated ($p<0.05$) in all three analyses. 5,703 SNPs were associated ($p<0.05$) in Single-SNP Analyses 1.1 and 2, but not Analysis 1.3. 11,291 SNPs were associated ($p<0.05$) in Single-SNP Analyses 1.1 and 1.3, but not Analysis 1.2. 743 SNPs were associated ($p<0.05$) in Single-SNP Analyses 1.2 and 1.3, but not Analysis 1.1. 9,321 SNPs were only associated ($p<0.05$) when all affected individuals were considered together, regardless of phenotypic subgroup assignment. 19,611 SNPs were uniquely associated ($p<0.05$) with individuals assigned to the 'less severe' phenotypic subgroup. 13,646 SNPs were uniquely associated ($p<0.05$) with individuals assigned to the 'more severe' phenotypic subgroup.

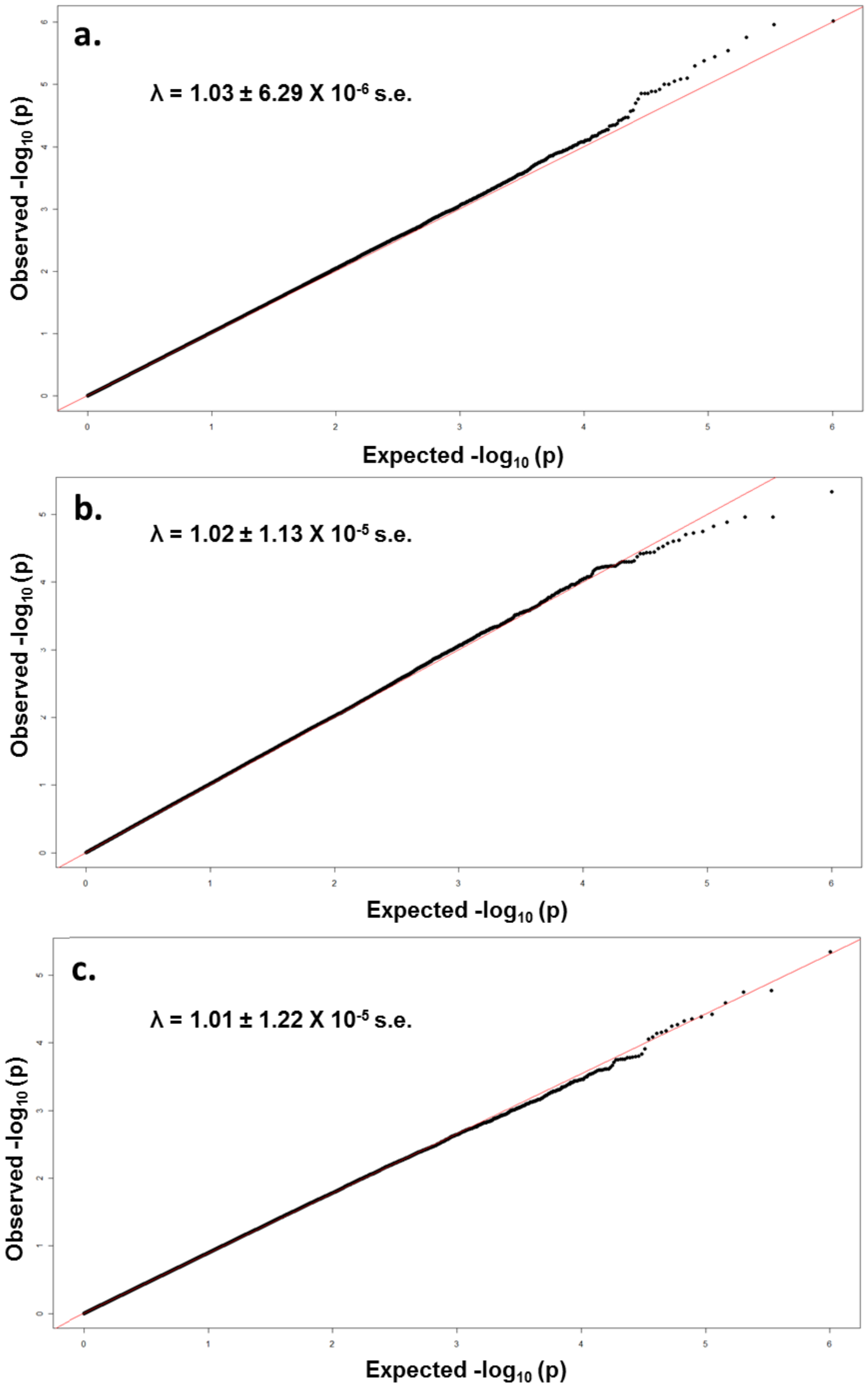


Figure 3.4. AGRE QQ Plots. Quantile-quantile plots of p-values from FBAT evaluating single-SNP associations with: **a.** all affected individuals **b.** 'less severe' subgroup **c.** 'more severe' subgroup. λ=genomic inflation factor; s.e.=standard error

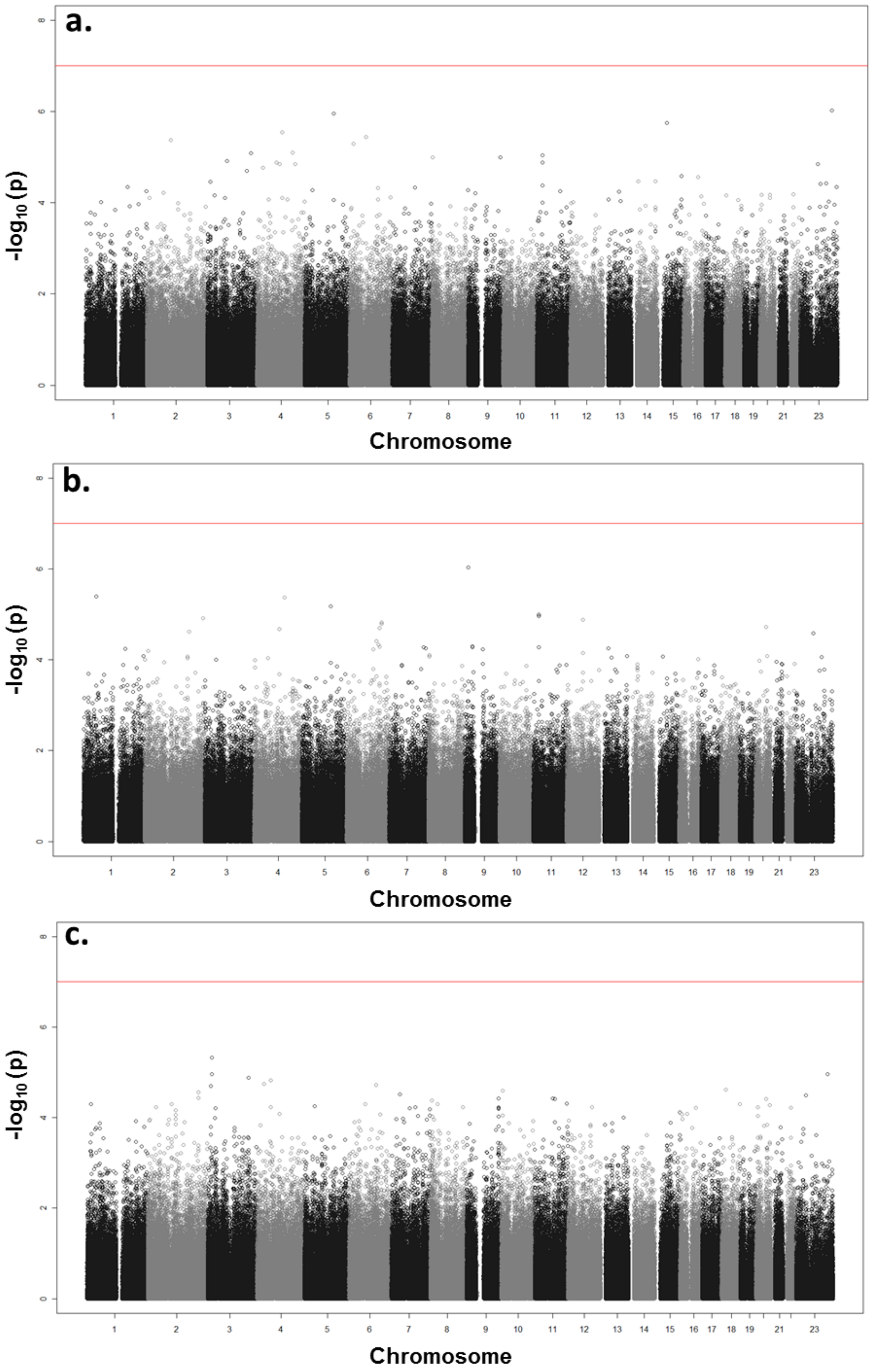


Figure 3.5. AGRE Genome-wide Single-SNP Association Results. Manhattan plots of p-values from FBAT evaluating SNP associations with: **a.** all affected individuals **b.** 'less severe' subgroup **c.** 'more severe' subgroup. Red line=Bonferroni corrected significance threshold ($p \leq 9.85 \times 10^{-8}$).

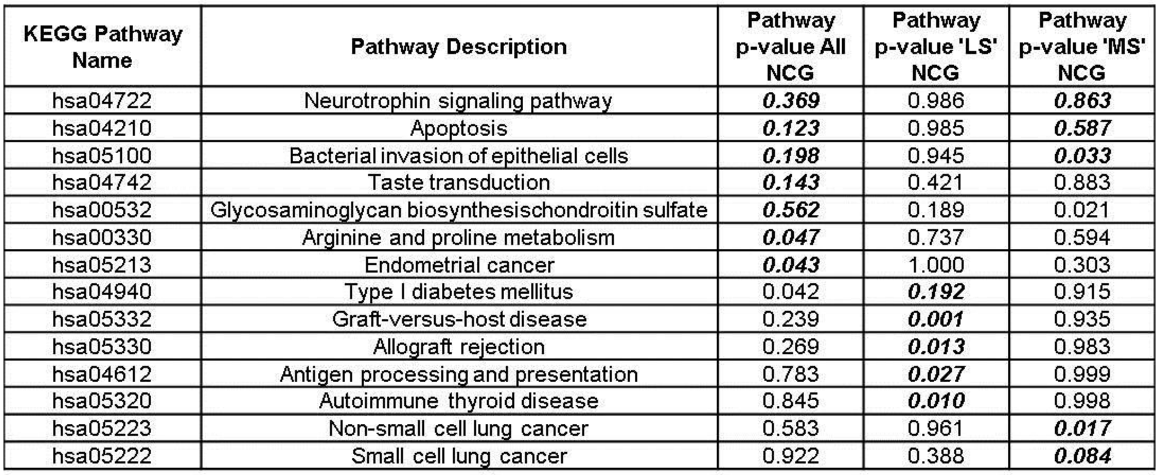
Pathway-Based Analyses: AGRE Dataset

We evaluated 209 pathways defined in the KEGG database for pathway-based association. We performed three separate pathway-based analyses using p-values generated via the three separate single-SNP analyses described above. We chose a threshold for significance at p<0.001 for pathway-based results. Seven KEGG pathways were associated (p<0.001) with in the full AGRE dataset. Three of these pathways remained associated when evaluating only the ‘more severe’ subgroup, while no pathways remained significant when evaluating only the ‘less severe’ subgroup (Table 3.3). Five KEGG pathways were exclusively associated (p<0.001) with cases in the ‘less severe’ subgroup (Table 3.3). Five different KEGG pathways were associated (p<0.001) with the ‘more severe’ subgroup. Two of these pathways were not associated in either of the other two pathway analyses (Table 3.3).

KEGG Pathway Name	Pathway Description	Pathway P-value All Affecteds	Pathway P-value "Less Severe"	Pathway P-value "More Severe"
hsa04722	Neurotrophin signaling pathway	< 0.001	0.800	< 0.001
hsa04210	Apoptosis	< 0.001	0.732	< 0.001
hsa05100	Bacterial invasion of epithelial cells	< 0.001	0.538	< 0.001
hsa04742	Taste transduction	< 0.001	0.018	0.124
hsa00532	Glycosaminoglycan biosynthesischondroitin sulfate	< 0.001	0.389	0.022
hsa00330	Arginine and proline metabolism	< 0.001	0.825	0.021
hsa05213	Endometrial cancer	< 0.001	0.997	0.008
hsa04940	Type I diabetes mellitus	0.082	< 0.001	0.375
hsa05332	Graft-versus-host disease	0.089	< 0.001	0.829
hsa05330	Allograft rejection	0.160	< 0.001	0.899
hsa04612	Antigen processing and presentation	0.241	< 0.001	0.975
hsa05320	Autoimmune thyroid disease	0.485	< 0.001	0.999
hsa05223	Non-small cell lung cancer	0.113	0.789	< 0.001
hsa05222	Small cell lung cancer	0.299	0.256	< 0.001

Table 3.3. Pathway-based Association Results: AGRE Dataset. Listed are biological pathways defined in the KEGG database that were associated (p<0.001) with at least one affection group. All=no phenotypic subgrouping; “LS”=individuals in the LS subgroup; “MS”=individuals in the MS subgroup.

Associated KEGG pathways were further subjected to permutation-based investigations to identify ‘core’ genes driving pathway associations. We defined ‘core’ genes as any gene whose p-value, in the context of the biological pathway, was p<0.001 and upon removal from analysis, the overall pathway p-value increased such that the previously implicated mechanism no longer met the significance threshold (Table 3.4). We identified 35 core genes within KEGG pathways associated (p<0.001) with all affected individuals, eight of these core genes function in ≥2 of these associated pathways. There are 39 genes total that associate (p<0.001) with all affected individuals, not all of these genes represent core genes driving pathway associations. Eleven of these genes did not meet our significance threshold for association in analyses where individuals were further defined by phenotypic subgroup (Table 3.5). We identified ten core genes within KEGG pathways associated (p<0.001) with the ‘less severe’ subgroup, five of these core genes function in ≥2 of these pathways. There are 18 total candidate genes associated (p<0.001) with the ‘less severe’ subgroup, eight of these genes did not meet our significance threshold when evaluating all affected individuals together, or the ‘more severe’ phenotypic subgroup (Table 3.5). We identified 24 core genes within KEGG pathways associated (p<0.001) with the ‘more severe’ subgroup, 10 of these genes function in ≥2 of these pathways. There are 34 total candidate genes associated (p<0.001) with the ‘more severe’ subgroup, 12 of these did not meet our significance threshold in any other analysis (Table 3.5).



KEGG Pathway Name	Pathway Description	Pathway p-value All NCG	Pathway p-value 'LS' NCG	Pathway p-value 'MS' NCG
hsa04722	Neurotrophin signaling pathway	***0.369***	0.986	***0.863***
hsa04210	Apoptosis	***0.123***	0.985	***0.587***
hsa05100	Bacterial invasion of epithelial cells	***0.198***	0.945	***0.033***
hsa04742	Taste transduction	***0.143***	0.421	0.883
hsa00532	Glycosaminoglycan biosynthesischondroitin sulfate	***0.562***	0.189	0.021
hsa00330	Arginine and proline metabolism	***0.047***	0.737	0.594
hsa05213	Endometrial cancer	***0.043***	1.000	0.303
hsa04940	Type I diabetes mellitus	0.042	***0.192***	0.915
hsa05332	Graft-versus-host disease	0.239	***0.001***	0.935
hsa05330	Allograft rejection	0.269	***0.013***	0.983
hsa04612	Antigen processing and presentation	0.783	***0.027***	0.999
hsa05320	Autoimmune thyroid disease	0.845	***0.010***	0.998
hsa05223	Non-small cell lung cancer	0.583	0.961	***0.017***
hsa05222	Small cell lung cancer	0.922	0.388	***0.084***

Table 3.4. Pathway-based Associations Following Removal of 'Core' Genes. Reported are p-values for biological pathways of interest following removal of SNPs assigned to suspected core genes. P-values in bold italics indicate these pathways met the significance threshold (p<0.001) in the full pathway-analysis for this affection group. All=no phenotypic subgrouping; 'LS'=individuals in the LS subgroup; 'MS'=individuals in the MS subgroup. NCG=no core genes included in analyses.

Encode ID	Location	Significant Pathways Gene Functions	Gene p-value All	Gene p-value 'LS'	Gene p-value 'MS'	OR All	95% CI		OR 'LS'	95% CI		OR 'MS'	95% CI	
NGF	1p13.1	hsa04722; hsa04210	< 0.001	0.594	< 0.001	1.21	1.19	1.24	NA	NA	NA	1.23	1.21	1.26
NRAS	1p13.2	hsa04722; hsa05223	< 0.001	1.000	< 0.001	1.16	1.07	1.25	NA	NA	NA	1.23	1.20	1.27
JUN	1p32-p31	hsa04722	0.003	0.018	< 0.001	NA	NA	NA	NA	NA	NA	1.45	1.37	1.52
CTSS	1q21	hsa04612	1.000	< 0.001	0.073	NA	NA	NA	1.25	1.24	1.27	NA	NA	NA
DNM3	1q24.3	hsa05100	< 0.001	0.592	< 0.001	1.22	1.20	1.24	NA	NA	NA	1.25	1.22	1.29
GLUL	1q31	hsa00330	< 0.001	1.000	< 0.001	1.23	1.17	1.30	NA	NA	NA	1.29	1.24	1.34
E2F1	20q11.2	hsa05223; hsa05222	< 0.001	< 0.001	< 0.001	1.18	1.17	1.18	1.42	0.97	2.08	1.25	1.06	1.48
KCNB1	20q13.2	hsa04742	< 0.001	0.094	< 0.001	1.20	1.18	1.22	NA	NA	NA	1.29	1.26	1.32
TPO	2p25	hsa05320	< 0.001	< 0.001	0.153	1.23	1.17	1.29	1.32	1.29	1.36	NA	NA	NA
PDK1	2q31.1	hsa04722	< 0.001	1.000	< 0.001	1.23	1.20	1.27	NA	NA	NA	1.30	1.23	1.37
FHIT	3p14.2	hsa05223; hsa05222	0.689	0.147	< 0.001	NA	NA	NA	NA	NA	NA	1.29	1.25	1.33
MLH1	3p21.3	hsa05213	< 0.001	1.000	1.000	1.16	1.02	1.31	NA	NA	NA	NA	NA	NA
RHOA	3p21.3	hsa05100; hsa04722	< 0.001	< 0.001	1.000	1.18	1.18	1.19	1.33	1.32	1.33	NA	NA	NA
RAF1	3p25	hsa04722; hsa05213	< 0.001	0.003	0.001	1.22	1.20	1.24	1.32	1.30	1.34	NA	NA	NA
CD80	3q13.3-q21	hsa05330; hsa04940; hsa05320; hsa05332	0.009	< 0.001	0.001	NA	NA	NA	1.36	1.32	1.39	NA	NA	NA
SGEF	3q25.2	hsa05100	0.001	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.30	1.25	1.35
CASP6	4q25	hsa04210	< 0.001	0.214	0.076	1.25	1.21	1.28	NA	NA	NA	NA	NA	NA
IL2	*4q26-q27*	*hsa04940; hsa05332; hsa05330; hsa05320*	*1.000*	*< 0.001*	*1.000*	*NA*	*NA*	*NA*	*1.23*	*0.96*	*1.56*	*NA*	*NA*	*NA*
TCF7	5q31.1	hsa05213	< 0.001	0.016	0.139	1.29	1.26	1.33	NA	NA	NA	NA	NA	NA
CTNNA1	5q31.2	hsa05213; hsa05100	< 0.001	1.000	< 0.001	1.15	1.13	1.16	NA	NA	NA	1.26	1.25	1.28
HLA-G	6p21.3	hsa04612; hsa04940; hsa05320; hsa05332; hsa05330	1.000	< 0.001	1.000	NA	NA	NA	1.58	1.45	1.70	NA	NA	NA
HLA-B	6p21.3	hsa05330; hsa04940; hsa05320; hsa05332	0.044	< 0.001	0.010	NA	NA	NA	1.31	1.27	1.34	NA	NA	NA
IL6	7p21	hsa05332	1.000	< 0.001	1.000	NA	NA	NA	1.44	1.39	1.49	NA	NA	NA
ASL	7q11.21	hsa00330	< 0.001	< 0.001	1.000	1.14	1.00	1.29	1.19	1.17	1.21	NA	NA	NA
WASL	7q31.3	hsa05100	< 0.001	1.000	1.000	1.21	1.04	1.42	NA	NA	NA	NA	NA	NA
TAS2R39	7q34	hsa04742	< 0.001	0.068	0.047	1.25	1.18	1.33	NA	NA	NA	NA	NA	NA
TAS2R40	7q34	hsa04742	< 0.001	1.000	0.014	1.29	1.15	1.44	NA	NA	NA	NA	NA	NA
TAS2R41	7q35	hsa04742	< 0.001	< 0.001	< 0.001	1.25	1.21	1.28	1.36	1.27	1.46	1.26	1.24	1.28
ARPC5L	9q33.3	hsa05100	1.000	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.31	1.13	1.48
ENDOG	9q34.1	hsa04210	< 0.001	< 0.001	1.000	1.20	1.17	1.22	1.48	1.44	1.51	NA	NA	NA
CBL	11q23.3	hsa05100	< 0.001	< 0.001	< 0.001	1.22	1.15	1.28	1.76	1.71	1.81	1.29	1.26	1.31
HSPA8	11q24.1	hsa04612	< 0.001	< 0.001	0.025	1.23	1.18	1.28	1.45	1.34	1.56	NA	NA	NA

Table 3.5. All Core Genes Driving Pathway-based Associations. Listed are core genes functioning in KEGG pathways associated at p<0.001 with at least one affection group in the AGRE analyses. Reported odds ratios (OR) represent an average for all SNPs assigned to the core gene boundary. 95% CI=confidence interval around mean OR for each gene. Genes in italics indicate that significance for this gene is based on one SNP assignment in pathway analyses. 95% CI represents the interval around the OR for these SNPs alone. All=no phenotypic subgrouping; 'LS'=individuals in the less severe subgroup; 'MS'=individuals in the more severe subgroup. NA indicates no odds ratios were calculated for SNPs in this analysis subgroup.

Encode ID	Location	Significant Pathways Gene Functions	Gene p-value All	Gene p-value 'LS'	Gene p-value 'MS'	OR All	95% CI		OR 'LS'	95% CI		OR 'MS'	95% CI	
TAS2R42	12p13	hsa04742	< 0.001	1.000	< 0.001	1.26	1.15	1.38	NA	NA	NA	1.28	1.02	1.59
TAS2R19	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
TAS2R20	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
TAS2R31	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
TAS2R43	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
TAS2R46	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
TAS2R50	*12p13.2*	*hsa04742*	*< 0.001*	*1.000*	*< 0.001*	*1.16*	*0.98*	*1.38*	*NA*	*NA*	*NA*	*1.28*	*1.03*	*1.60*
IRAK4	12q12	hsa04722; hsa04210	0.008	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.30	1.19	1.41
GLS2	12q13	hsa00330	< 0.001	< 0.001	0.015	1.16	1.13	1.19	1.26	1.20	1.32	NA	NA	NA
CDK4	12q14	hsa05223; hsa05222	0.065	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.26	1.18	1.34
APAF1	12q23	hsa04210	0.007	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.25	1.19	1.31
ARPC3	12q24.11	hsa05100	1.000	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.26	1.23	1.29
RB1	13q14.2	hsa05223; hsa05222	< 0.001	1.000	< 0.001	1.27	1.23	1.30	NA	NA	NA	1.34	1.11	1.62
NFKBIA	14q13	hsa04210	0.038	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.30	1.21	1.39
MAP2K5	15q23	hsa04722	0.010	0.176	< 0.001	NA	NA	NA	NA	NA	NA	1.28	1.23	1.32
NTRK3	15q25	hsa04722	0.055	0.597	< 0.001	NA	NA	NA	NA	NA	NA	1.23	1.21	1.25
GOT2	16q21	hsa00330	< 0.001	1.000	0.007	1.34	1.30	1.39	NA	NA	NA	NA	NA	NA
PIK3R5	17p13.1	hsa05213; hsa05100; hsa04722; hsa04210	< 0.001	0.004	0.090	1.21	1.18	1.23	NA	NA	NA	NA	NA	NA
NOS2	17q11.2-q12	hsa00330	< 0.001	0.137	< 0.001	1.23	1.19	1.28	NA	NA	NA	1.28	1.25	1.32
PIK3R2	19q13.2-q13.4	hsa05223; hsa05222; hsa05100; hsa04722; hsa04210	0.029	1.000	< 0.001	NA	NA	NA	NA	NA	NA	1.24	1.22	1.27
BAX	19q13.3-q13.4	hsa04210; hsa4722	< 0.001	1.000	0.016	1.20	1.19	1.21	NA	NA	NA	NA	NA	NA
KIR2DL3	19q13.4	hsa04612; hsa05332	0.034	< 0.001	0.196	NA	NA	NA	1.30	1.26	1.35	NA	NA	NA
KIR3DL3	19q13.42	hsa04612	0.008	< 0.001	0.210	NA	NA	NA	1.30	1.27	1.34	NA	NA	NA
RPS6KA3	Xp22.2-p22.1	hsa04722	< 0.001	0.020	0.173	1.94	1.53	2.34	NA	NA	NA	NA	NA	NA
IRS4	Xq22.3	hsa04722	< 0.001	1.000	< 0.001	1.47	1.32	1.63	NA	NA	NA	1.69	1.52	1.86
IKBKG	Xq28	hsa04210	< 0.001	< 0.001	< 0.001	2.86	1.80	4.55	2.56	1.21	5.38	3.06	1.74	5.38
IRAK1	Xq28	hsa04722; hsa04210	< 0.001	1.000	< 0.001	1.85	1.16	2.54	NA	NA	NA	1.95	1.69	2.20

Table 3.5 (CONTINUED). All Core Genes Driving Pathway-based Associations.

Single-SNP Association Analyses: AGP Dataset

A total of 779,343 SNPs were analyzed for association in the validation analyses. These SNPs were evaluated for association with all AGP dataset individuals meeting diagnostic criteria for an ASD on both the ADI-R and ADOS (Single-SNP Analysis 2.1), only affected individuals assigned to the ‘less severe’ phenotypic subgroup (Single-SNP Analysis 2.2), and only affected individuals assigned to the ‘more severe’ ASD subgroup (Single-SNP Analysis 2.3) (Figure 3.3). Genomic inflation factors for these analyses were 1.028, 1.017, and 1.017, respectively (Figure 3.6). Nine SNPs met a Bonferroni corrected significance threshold of $p \leq 6.42 \times 10^{-8}$ in Single-SNP Analysis 2.1 (Figure 3.7). Associations for these markers have not previously been reported as the sex chromosomes were not included in these analyses[134]. However, the current version of FBAT allows for evaluation of markers on the sex chromosomes (http://www.biostat.harvard.edu/fbat/fbat.htm). Further information on SNPs surpassing a Bonferroni corrected significance threshold is provided in Table 3.6 and Figure 3.8.

From Single-SNP Analysis 2.1, there were 41,331 SNPs ($p<0.05$) that were evaluated in Pathway Analysis 2.1. From Single-SNP Analysis 2.2, 40,953 SNPs were evaluated in Pathway Analysis 2.2. From Single-SNP Analysis 2.3, 40,375 SNPs were evaluated in Pathway Analysis 2.3. Only 1,140 SNPs were associated ($p<0.05$) in all three analyses. 10,531 SNPs were associated ($p<0.05$) in Single-SNP Analyses 2.1 and 2.2, but not Analysis 2.3. 15,730 SNPs were associated ($p<0.05$) in Single-SNP Analyses 2.1 and 2.3, but not Analysis 2.2. 1,016 SNPs were associated ($p<0.05$) in Single-SNP Analyses 2.2 and 2.3, but not Analysis 2.1. 13,930 SNPs were associated ($p<0.05$) only when all affected individuals were considered together, regardless of phenotypic subgroup assignment. 28,266 SNPs were uniquely associated ($p<0.05$) with individuals assigned to the ‘less severe’ phenotypic subgroup. 22,489 SNPs were uniquely associated (p<0.05) with individuals assigned to the ‘more severe’ phenotypic subgroup.

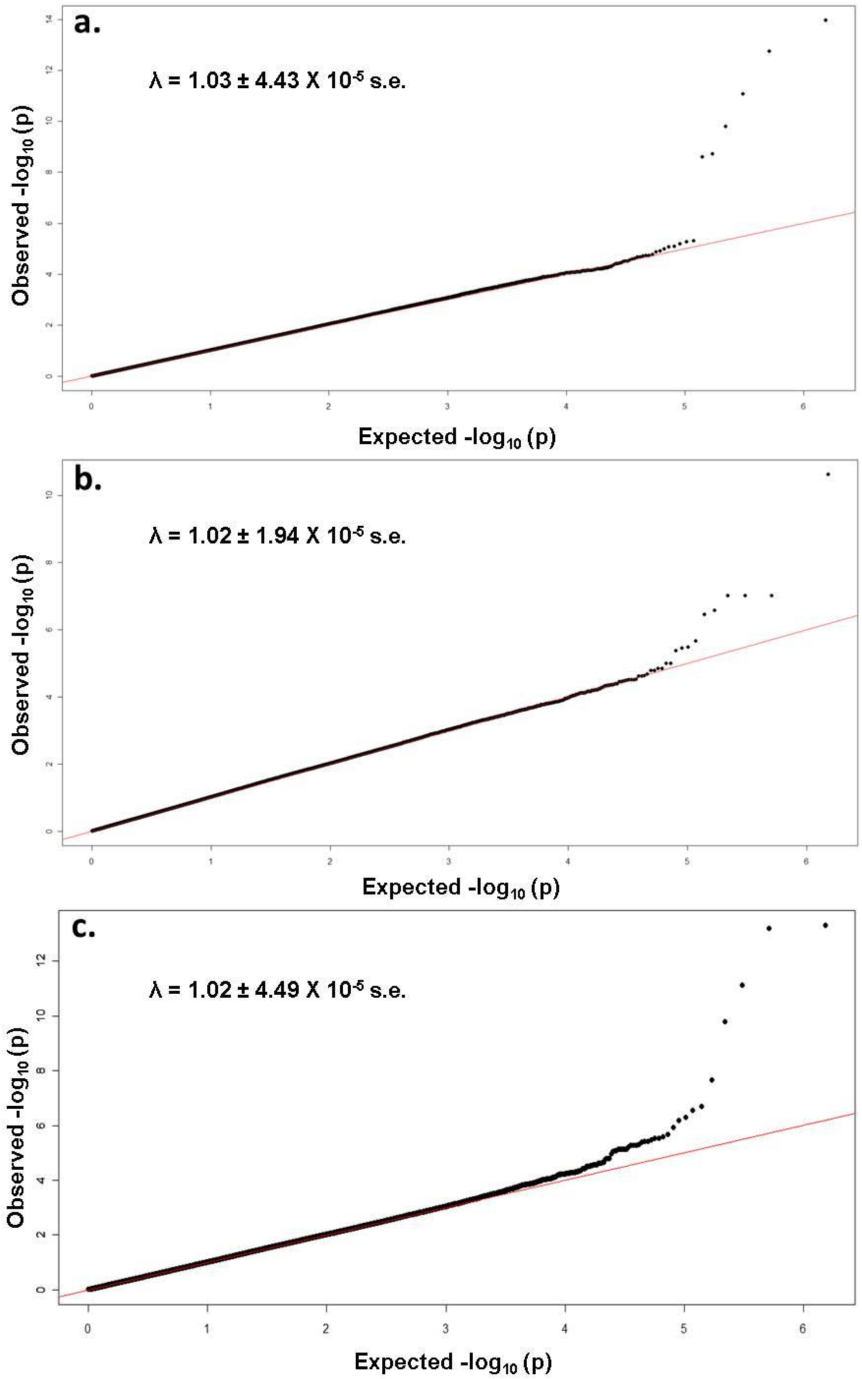


Figure 3.6. AGP QQ Plots. Quantile-quantile plots of p-values from FBAT evaluating single-SNP associations with: **a.** all affected individuals **b.** 'less severe' subgroup **c.** 'more severe' subgroup. λ=genomic inflation factor; s.e.=standard error.

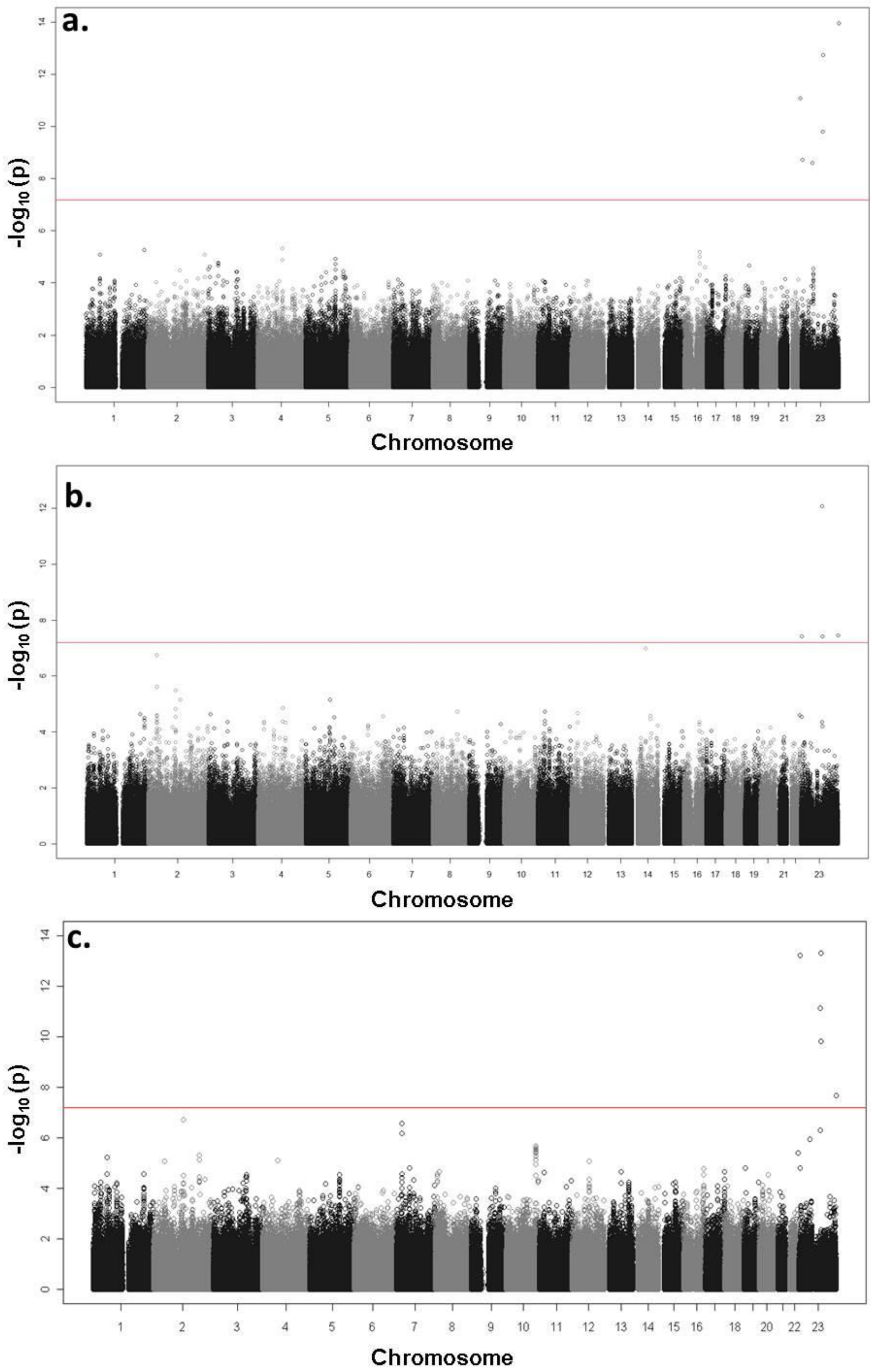


Figure 3.7. AGP Genome-wide Single-SNP Association Results. Manhattan plots of p-values from FBAT evaluating SNP associations with: **a.** all affected individuals **b.** 'less severe' subgroup **c.** 'more severe' subgoupr. Red line=Bonferroni corrected significance threshold ($p \leq 6.42 \times 10^{-8}$).

SNP	Location	All Affecteds p-value	Male Only p-value	Female Only p-value	'Less Severe' p-value	'More Severe' p-value	Mat. to Males (U:T)	Mat. to Females (U:T)	MAF Overall	HWE	Nearby Genes (+/- 50kb)
rs2896799	Xp22.31	0.00E+00	1.67E-15	1.70E-01	9.61E-08	6.07E-14	3:61	16:25	8.85E-02	8.00E-01	KAL1 VCX3B
rs34537684	Xq21.31; Yp11.2	0.00E+00	2.11E-15	4.50E-01	2.17E-06	4.77E-14	0:63	17:13	9.27E-02	3.36E-01	KRT18P11 SNX3P1X; SNX3P1Y PCDH11X; PCDH11Y
rs6652550	Xq21.31; Yp11.2	0.00E+00	1.22E-15	4.50E-01	2.36E-11	7.55E-12	0:64	12:15	8.53E-02	7.97E-01	none
rs909439	Xq28; Yq12	1.08E-14	6.80E-08	1.34E-01	9.43E-08	2.09E-08	1:32	5:11	5.18E-02	9.92E-02	VAMP7
rs34013457	Xq21.31; Yp11.2	1.78E-13	9.24E-13	1.44E-01	1.83E-04	1.55E-10	0:51	20:12	6.74E-02	6.78E-01	KRT18P11 PCDH11X; PCDH11Y
rs4074620	Xp22.33; Yp11.3	8.30E-12	8.72E-09	6.17E-01	3.46E-07	3.86E-06	2:36	9:06	5.52E-02	3.59E-01	PLCXD1 GTPBP6
rs5941356	Xq21.31	1.59E-10	6.98E-08	4.65E-01	6.30E-05	5.03E-07	18:62	17:13	9.54E-02	6.66E-01	none
rs1183735	Xp22.31	1.87E-09	1.45E-06	1.43E-02	2.90E-05	1.60E-05	17:53	18:06	8.58E-02	2.32E-03	XR_110926.1
rs5906541	Xp11.23	2.43E-09	4.89E-10	2.73E-01	5.56E-04	1.14E-06	3:42	18:12	5.10E-02	7.64E-04	SSX6

Table 3.6. Details for Single-SNP Associations Passing a Bonferroni-corrected Significance Threshold. Listed are SNPs whose p-value from FBAT was less than a Bonferroni-corrected significance threshold ($p \leq 9.85 \times 10^{-8}$). Reported are the number of informative transmissions for all SNPs from the Maternal (Mat.) lineage. U:T indicates the untransmitted to transmitted ratio for risk alleles at these markers. MAF=overall minor allele frequency for the entire AGP dataset; HWE=p-value for Hardy-Weinberg equilibrium.

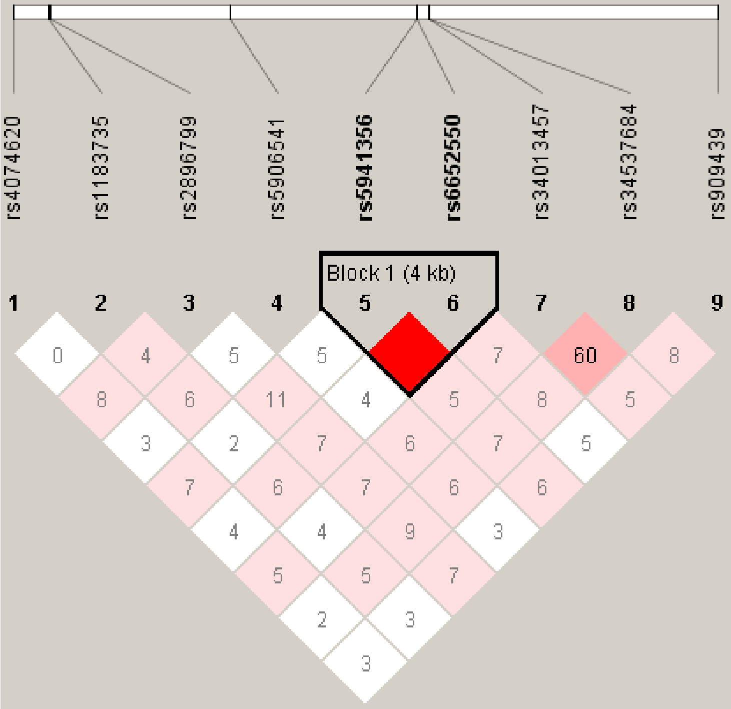


Figure 3.8. Linkage Disequilibrium Structure for Very Significant Single-SNP Associations. Haploview plot of D' values calculated for X chromosome SNPs surpassing a Bonferroni-adjusted significance threshold for association in the AGP dataset single-SNP analyses. A predicted haplotype block, including markers rs5941356, & rs6652550, is indicated (D'=1.0, r^2=0.012).

Pathway-Based Analyses: AGP Dataset

To determine which AGRE associated KEGG pathways validated in the AGP dataset, we chose a threshold for pathway significance at p<0.05. A total of seven pathways validated in the AGP dataset at this significance threshold. The pathway defined in KEGG as 'Bacterial invasion of epithelial cells' validated not only across datasets, but was associated with the 'more severe' subgroups from both datasets (Table 3.7; Figure 3.9). The other six pathways that validate in the AGP dataset are associated independent of phenotypic subgroup assignment. For example, the pathway defined as 'Allograft rejection' is associated (p<0.001) with the 'less severe' AGRE subgroup and the 'more severe' AGP subgroup (Table 3.7). There are another 13 KEGG pathways that are trending towards significance (p<0.05) in at least one analysis for both datasets (Table 3.8). We further investigated validated pathways to identify genes driving pathway associations and compared these results with core genes identified with the AGRE dataset (Table 3.9). Four core genes identified in the 'less severe' AGRE subgroup analysis validated (p<0.001) in the 'less severe' AGP subgroup analysis, and five core genes identified in the 'more severe' AGRE subgroup analysis validated (p<0.001) in the 'more severe' AGP subgroup analysis (Table 3.9). In some cases, the same specific gene did not validate but genes within the same gene family were identified as driving pathway associations in both datasets. For example, the *ARPC3* and *ARPC5* genes are significantly associated (p<0.001) with the 'more severe' AGRE subgroup while the *ARPC1A* gene is significantly associated (p<0.001) with the 'more severe' AGP subgroup (Figure 3.9).

KEGG Pathway ID	Pathway Description	All p-value (AGRE)	All p-value (AGP)	'LS' p-value (AGRE)	'LS' p-value (AGP)	'MS' p-value (AGRE)	'MS' p-value (AGP)
hsa05100	***Bacterial invasion of epithelial cells***	***< 0.001***	***0.002***	***0.538***	***0.689***	***< 0.001***	***0.021***
hsa05330	Allograft rejection	0.160	**< 0.001**	**< 0.001**	1.000	0.899	**< 0.001**
hsa04940	Type I diabetes mellitus	0.082	**< 0.001**	**< 0.001**	1.000	0.375	**< 0.001**
hsa05332	Graft-versus-host disease	0.089	**< 0.001**	**< 0.001**	1.000	0.829	**< 0.001**
hsa04612	Antigen processing and presentation	0.241	**< 0.001**	**< 0.001**	1.000	0.975	**< 0.001**
hsa05320	Autoimmune thyroid disease	0.485	**< 0.001**	**< 0.001**	1.000	0.999	**< 0.001**
hsa04722	Neurotrophin signaling pathway	**< 0.001**	0.168	0.800	**0.003**	**< 0.001**	0.106

Table 3.7. Validated Pathway-based Association Results. Listed are biological pathways defined in the KEGG database that validated (p<0.05) in at least one affection group in the AGP dataset. All=no phenotypic subgrouping; 'LS'=individuals in the 'less severe' subgroups; 'MS'=individuals in the 'more severe' subgroups.

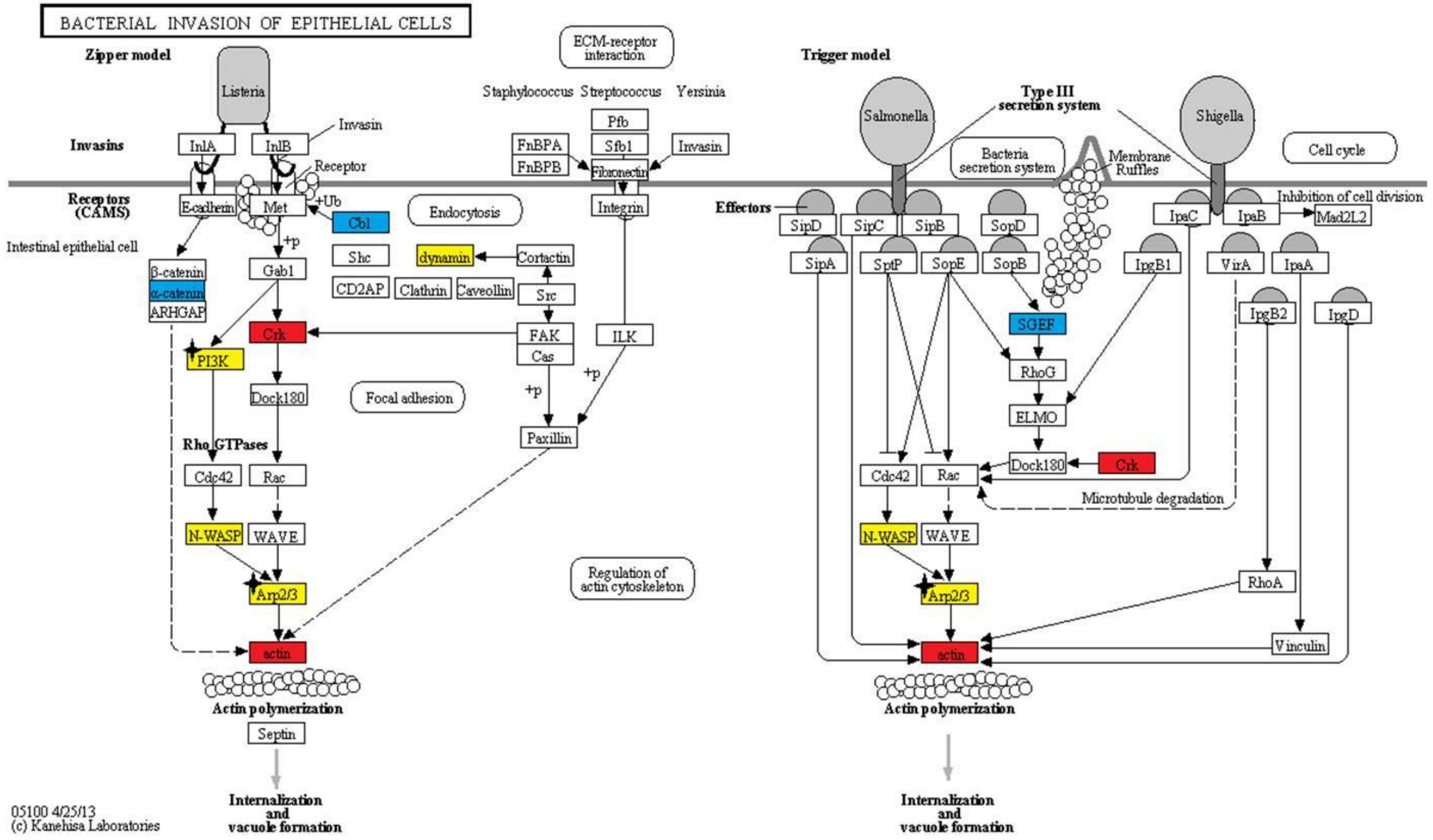


Figure 3.9. Core Genes in Top Validated Pathway. Shown is KEGG detail regarding associated genes functioning in the 'Bacterial invasion of epithelial cells' pathway. This pathway was significantly associated (p<0.001) with the 'more severe' subgroup in the AGRE dataset and was also associated (p=0.021) with the 'more severe' subgroup in the AGP dataset. Associated genes (p<0.001) are color-coded; blue=AGRE only; red=AGP only; yellow=Both datasets. Starred genes indicate that different genes within the same gene family were uniquely associated in each dataset.

KEGG Pathway Name	Pathway Description	Pathway p-value All Affecteds (AGRE)	Pathway p-value All Affecteds (AGP)	Pathway p-value 'Less Severe' (AGRE)	Pathway p-value 'Less Severe' (AGP)	Pathway p-value 'More Severe' (AGRE)	Pathway p-value 'More Severe' (AGP)
hsa05100	***Bacterial invasion of epithelial cells***	***< 0.001***	***0.002***	***0.538***	***0.689***	***< 0.001***	***0.021***
hsa05330	**Allograft rejection**	0.160	**< 0.001**	**< 0.001**	1.000	0.899	**< 0.001**
hsa04940	**Type I diabetes mellitus**	0.082	**< 0.001**	**< 0.001**	1.000	0.375	**< 0.001**
hsa05332	**Graft-versus-host disease**	0.089	**< 0.001**	**< 0.001**	1.000	0.829	**< 0.001**
hsa04612	**Antigen processing and presentation**	0.241	**< 0.001**	**< 0.001**	1.000	0.975	**< 0.001**
hsa05320	**Autoimmune thyroid disease**	0.485	**< 0.001**	**< 0.001**	1.000	0.999	**< 0.001**
hsa04722	**Neurotrophin signaling pathway**	**< 0.001**	0.168	0.800	**0.003**	**< 0.001**	0.106
hsa05210	**Colorectal cancer**	**0.001**	**0.021**	0.226	0.378	**0.009**	**< 0.001**
hsa00360	**Phenylalanine metabolism**	**0.005**	0.057	0.125	**< 0.001**	**0.007**	**0.009**
hsa05211	**Renal cell carcinoma**	**0.008**	0.181	0.096	**< 0.001**	**0.027**	**0.015**
hsa00592	**alpha-Linolenic acid metabolism**	**0.039**	**< 0.001**	0.338	0.472	0.179	**< 0.001**
hsa05142	**Chagas disease**	**0.039**	**0.004**	0.502	**< 0.001**	**0.001**	0.146
hsa00400	**Phenylalanine, tyrosine and tryptophan biosynthesis**	**0.047**	0.121	0.537	**< 0.001**	**0.015**	0.188
hsa05416	**Viral myocarditis**	0.096	**< 0.001**	**0.001**	0.998	0.235	**< 0.001**
hsa00250	**Alanine, aspartate and glutamate metabolism**	0.168	**< 0.001**	0.828	**0.002**	**0.030**	**< 0.001**
hsa05212	**Pancreatic cancer**	0.307	0.285	0.542	0.474	**0.028**	**< 0.001**
hsa04012	**ErbB signaling pathway**	0.354	0.097	0.482	0.173	**0.014**	**< 0.001**
hsa03060	**Protein export**	0.463	0.132	0.545	0.238	**0.025**	**< 0.001**
hsa04672	**Intestinal immune network for IgA production**	0.558	0.159	**0.016**	0.960	0.996	**< 0.001**
hsa00591	**Linoleic acid metabolism**	0.846	**< 0.001**	**0.016**	**0.032**	0.879	**< 0.001**
hsa04210	Apoptosis	< 0.001	0.080	0.732	0.057	< 0.001	0.154
hsa05213	Endometrial cancer	< 0.001	0.574	0.997	0.837	0.008	0.084
hsa00330	Arginine and proline metabolism	< 0.001	0.863	0.825	0.958	0.021	0.884
hsa00532	Glycosaminoglycan biosynthesis - chondroitin sulfate	< 0.001	0.923	0.389	0.827	0.022	0.998
hsa04742	Taste transduction	< 0.001	1.000	0.018	0.999	0.124	0.539
hsa04664	Fc epsilon RI signaling pathway	0.103	< 0.001	0.860	0.132	0.079	< 0.001
hsa05131	Shigellosis	0.834	< 0.001	0.234	0.090	0.099	< 0.001
hsa04914	Progesterone-mediated oocyte maturation	0.011	0.018	0.159	< 0.001	0.453	0.107
hsa05010	Alzheimer's disease	0.994	0.173	0.953	< 0.001	0.505	0.615
hsa00562	Inositol phosphate metabolism	1.000	0.001	0.997	< 0.001	0.725	0.046
hsa05223	Non-small cell lung cancer	0.113	0.430	0.789	0.053	< 0.001	0.081
hsa05222	Small cell lung cancer	0.299	0.362	0.256	0.056	< 0.001	0.344
hsa04666	Fc gamma R-mediated phagocytosis	0.156	0.170	0.621	0.417	0.232	< 0.001
hsa04640	Hematopoietic cell lineage	0.472	0.113	0.060	0.999	0.847	< 0.001
hsa05310	Asthma	0.687	0.614	0.076	1.000	1.000	< 0.001

Table 3.8. Pathway-based Association Results: Both Datasets. Listed are biological pathways defined in the KEGG database that associated at p<0.001 with at least one affection group in either the AGRE or AGP dataset. Pathways indicated in bold validate (p<0.05) across datasets. All Affecteds=no phenotypic subgrouping; 'Less Severe'=individuals in the 'less severe' subgroup; 'More Severe'=individuals in the 'more severe' subgroup.

Encode ID	Location	Significant Pathways Gene Functions	Gene p-value All	OR All	95% CI		Gene p-value 'LS'	OR 'LS'	95% CI		Gene p-value 'MS'	OR 'MS'	95% CI	
NGF	1p13.1	hsa04722; hsa04210	0.134	NA	NA	NA	0.612	NA	NA	NA	0.145	NA	NA	NA
NRAS	1p13.2	hsa04722; hsa05223	0.142	NA	NA	NA	0.122	NA	NA	NA	1.000	NA	NA	NA
JUN	1p32-p31	hsa04722	1.000	NA	NA	NA	**0.038**	**1.26**	**1.15**	**1.36**	1.000	NA	NA	NA
CTSS	1q21	hsa04612	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
DNM3	1q24.3	hsa05100	**0.023**	**1.16**	**1.13**	**1.20**	0.126	NA	NA	NA	**0.016**	**1.23**	**1.19**	**1.28**
GLUL	1q31	hsa00330	0.079	NA	NA	NA	1.000	NA	NA	NA	**0.011**	**1.20**	**1.16**	**1.25**
TPO	2p25	hsa05320	**0.007**	**1.16**	**1.12**	**1.21**	0.497	NA	NA	NA	0.497	NA	NA	NA
PDK1	2q31.1	hsa04722	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
FHIT	3p14.2	hsa05223; hsa05222	0.079	NA	NA	NA	0.768	NA	NA	NA	0.082	NA	NA	NA
RHOA	3p21.3	hsa04722; hsa5100	1.000	NA	NA	NA	**< 0.001**	**1.36**	**1.14**	**1.58**	1.000	NA	NA	NA
MLH1	3p21.3	hsa05213	1.000	NA	NA	NA	1.000	NA	NA	NA	**< 0.001**	**1.20**	**1.15**	**1.25**
RAF1	3p25	hsa04722; hsa05213	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
CD80	3q13.3-q21	hsa05330; hsa04940; hsa05320; hsa05332	**0.002**	**1.20**	**1.16**	**1.23**	1.000	NA	NA	NA	**< 0.001**	**1.25**	**1.18**	**1.32**
SGEF	3q25.2	hsa05100	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
CASP6	4q25	hsa04210	0.287	NA	NA	NA	0.080	NA	NA	NA	0.068	NA	NA	NA
IL2	*4q26-q27*	*hsa04940; hsa05332; hsa05330; hsa05320*	*1.000*	*NA*	*NA*	*NA*	***< 0.001***	***1.23***	***1.02***	***1.47***	*1.000*	*NA*	*NA*	*NA*
TCF7	5q31.1	hsa05213	1.000	NA	NA	NA	1.000	NA	NA	NA	0.192	NA	NA	NA
CTNNA1	5q31.2	hsa05213; hsa05100	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
HLA-G	6p21.3	hsa04612; hsa04940; hsa05320; hsa05332; hsa05330	**< 0.001**	**1.15**	**1.14**	**1.16**	1.000	NA	NA	NA	**0.009**	**1.21**	**1.17**	**1.25**
HLA-B	6p21.3	hsa05330; hsa04940; hsa05320; hsa05332	0.211	NA	NA	NA	0.431	NA	NA	NA	**0.005**	**1.22**	**1.14**	**1.30**
IL6	7p21	hsa05332	0.211	NA	NA	NA	1.000	NA	NA	NA	0.207	NA	NA	NA
ASL	7q11.21	hsa00330	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
WASL	7q31.3	hsa05100	**< 0.001**	**1.11**	**1.11**	**1.12**	**< 0.001**	**1.25**	**1.22**	**1.27**	1.000	NA	NA	NA
TAS2R39	7q34	hsa04742	0.102	NA	NA	NA	0.098	NA	NA	NA	**0.009**	**1.20**	**1.14**	**1.26**
TAS2R40	7q34	hsa04742	1.000	NA	NA	NA	0.057	NA	NA	NA	**0.002**	**1.22**	**1.14**	**1.31**
TAS2R41	7q35	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
ARPC5L	9q33.3	hsa05100	**0.033**	**1.18**	**1.15**	**1.21**	**0.020**	**1.25**	**1.21**	**1.29**	**0.042**	**1.22**	**1.19**	**1.25**
ENDOG	9q34.1	hsa04210	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
CBL	11q23.3	hsa05100	1.000	NA	NA	NA	1.000	NA	NA	NA	0.105	NA	NA	NA
HSPA8	11q24.1	hsa04612	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA

Table 3.9. Validation Dataset P-values and Odds Ratios Calculations for 'Core Genes' Identified in the Discovery Dataset. Listed are statistics calculated in the AGP dataset for genes of interest identified in the AGRE dataset. These genes function in validated KEGG pathways. Validated gene associations are indicated in bold. Reported odds ratios (OR) represent an average for all SNPs assigned to the core gene boundary. 95% C.I. represents the confidence interval around the mean odds ratio for each gene. All=no phenotypic subgrouping; 'LS'=individuals in the less severe subgroup; 'MS'=individuals in the more severe subgroup. NA indicates no ORs were calculated for SNPs in this analysis subgroup.

Encode ID	Location	Significant Pathways Gene Functions	Gene p-value All	OR All	95% CI		Gene p-value 'LS'	OR 'LS'	95% CI		Gene p-value 'MS'	OR 'MS'	95% CI	
TAS2R42	12p13	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R43	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R46	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R31	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R19	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R20	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
TAS2R50	12p13.2	hsa04742	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
IRAK4	12q12	hsa04722; hsa04210	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
GLS2	12q13	hsa00330	1.000	NA	NA	NA	**0.008**	**1.38**	**0.99**	**1.77**	1.000	NA	NA	NA
CDK4	12q14	hsa05223; hsa05222	0.097	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
APAF1	12q23	hsa04210	**< 0.001**	**1.13**	**1.13**	**1.14**	1.000	NA	NA	NA	**< 0.001**	**1.26**	**1.21**	**1.32**
ARPC3	12q24.11	hsa05100	**< 0.001**	**1.15**	**1.00**	**1.32**	1.000	NA	NA	NA	1.000	NA	NA	NA
RB1	13q14.2	hsa05223; hsa05222	1.000	NA	NA	NA	1.000	NA	NA	NA	**< 0.001**	**1.15**	**1.00**	**1.33**
NFKBIA	14q13	hsa04210	**0.041**	**1.13**	**1.11**	**1.14**	0.195	NA	NA	NA	0.192	NA	NA	NA
MAP2K5	15q23	hsa04722	1.000	NA	NA	NA	0.501	NA	NA	NA	1.000	NA	NA	NA
NTRK3	15q25	hsa04722	1.000	NA	NA	NA	0.977	NA	NA	NA	0.863	NA	NA	NA
GOT2	16q21	hsa00330	1.000	NA	NA	NA	0.188	NA	NA	NA	1.000	NA	NA	NA
PIK3R5	17p13.1	hsa05213; hsa05100; hsa04722; hsa04210	0.384	NA	NA	NA	**0.023**	**1.44**	**1.23**	**1.65**	0.130	NA	NA	NA
NOS2	17q11.2-q12	hsa00330	0.257	NA	NA	NA	0.227	NA	NA	NA	0.218	NA	NA	NA
PIK3R2	19q13.2-q13.4	hsa05223; hsa05222; hsa05100; hsa04722; hsa04210	**< 0.001**	**1.13**	**1.12**	**1.14**	1.000	NA	NA	NA	**< 0.001**	**1.31**	**1.22**	**1.39**
BAX	19q13.3-q13.4	hsa04210; hsa4722	1.000	NA	NA	NA	0.054	NA	NA	NA	1.000	NA	NA	NA
KIR2DL3	19q13.4	hsa04612; hsa05332	**0.043**	**1.14**	**1.03**	**1.25**	**0.001**	**1.29**	**1.24**	**1.33**	1.000	NA	NA	NA
KIR3DL3	19q13.42	hsa04612	0.092	NA	NA	NA	**< 0.001**	**1.29**	**1.24**	**1.33**	1.000	NA	NA	NA
E2F1	20q11.2	hsa05223; hsa05222	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
KCNB1	20q13.2	hsa04742	0.080	NA	NA	NA	1.000	NA	NA	NA	**< 0.001**	**1.26**	**1.20**	**1.32**
RPS6KA3	Xp22.2-p22.1	hsa04722	1.000	NA	NA	NA	1.000	NA	NA	NA	1.000	NA	NA	NA
IRS4	Xq22.3	hsa04722	1.000	NA	NA	NA	**< 0.001**	**1.39**	**1.24**	**1.54**	**< 0.001**	**1.35**	**1.21**	**1.48**
IRAK1	Xq28	hsa04722; hsa04210	1.000	NA	NA	NA	**< 0.001**	**1.52**	**1.11**	**2.08**	1.000	NA	NA	NA
IKBKG	Xq28	hsa04210	**< 0.001**	**1.28**	**1.05**	**1.57**	**< 0.001**	**1.36**	**0.56**	**0.96**	1.000	NA	NA	NA

Table 3.9. (CONTINUED). Validation Dataset P-values and Odds Ratios Calculations for 'Core Genes' Identified in the Discovery Dataset.

Discussion

Our results show that different genetic factors are uniquely associated with ASD subgroups defined using multiple sources of behavioral and clinical exam information. This suggests that meaningful phenotypic subgroup definitions can help clarify the underlying genetic etiology of ASD. Further, the pathway-based approach seemed to be a more biologically relevant way to evaluate the risk effects of common, inherited variation, as opposed to single-variant analysis. The vast majority of the SNPs evaluated did not meet a multiple-testing adjusted significance threshold when analyzed individually. However, by evaluating the combined effects from many SNPs, we were able to identify groups of genes with similar function contributing to risk for ASD and effectively account for underlying genetic heterogeneity across two independent ASD datasets. By using the combined approach of phenotypic subgrouping and pathway-based genetic analysis, we were able to implicate functional pathways of interest and refine the genetic bottlenecks related to specific ASD traits.

By subgrouping individuals based on similar expression of ASD-related phenotypes, we drastically reduced the number of cases evaluated in subgroup-specific analyses (AGRE=35%-65% reduction; AGP=40%-60% reduction). Despite these substantial reductions in sample size, subgroup-specific odds ratio (OR) calculations for core genes that were also associated when analyzing all cases together indicate no reduction in observed genetic effects. In fact, the effects on risk only seem to increase in subgroup-specific analyses. This suggested our method of phenotypic subgrouping potentially reduced statistical noise and increased the ability to detect genetic effects.

Performing phenotypic subgroup-specific genetic analyses also allowed us to more easily refine potential phenotype-genotype relationships. For example, we observed that pathways and 'core' genes related to adaptive immunity were almost exclusively associated with the ‘less severe’ AGRE subgroup. Two 'core' genes in all of the pathways associated with the ‘less severe’ AGRE subgroup were located in the major histocompatibility complex (*HLA-B*, *HLA-G*). Increasing evidence supports substantial neural-immune crosstalk starting in the fetal brain and continuing throughout life[38, 145]. Many members of the major histocompatibility complex are thought to play important roles in brain development and function, reviewed in Needleman & McAllister, 2012[153]. Multiple studies have also identified extensive changes in the immune systems of individuals with ASD, reviewed in Careaga, 2010[37]. Interestingly, two pathways significantly associated with the ‘less severe’ AGRE subgroup are ‘autoimmune thyroid disease’ and ‘type-I diabetes mellitus’. A few epidemiological studies have reported associations between both of these diseases and ASD[19, 149, 150]. Specifically, autoimmune thyroid disease is more frequent in children diagnosed with a regressive form of ASD, compared to children diagnosed with an early-onset form[149]. It is assumed that cases with regressive ASD exhibit less delayed early development[100]. We saw that the ‘abnormality of development evident ≤ 36 months’ domain score from the ADI-R stood out as having a strong influence on assignment of individuals to our ASD subgroups. Higher scores on this measure indicate very severe abnormality of development observed early in life. All individuals assigned to the ‘less severe’ AGRE subgroup had low severity scores on this measure; it is possible that some of these individuals exhibited a regressive form of ASD. Cases in the AGP subgroups were older, on average, at the time of ADI-R than were cases in the AGRE subgroups ($t_{MoreSevere}$=5.01, p<0.00001; $t_{LessSevere}$=2.10, p=0.017). A larger portion of individuals in the ‘more severe’ AGP subgroup have less severe scores on the 'abnormality of development evident ≤ 36 months’ measure compared to individuals in the ‘more severe’ AGRE subgroup ($z_{MannWhitney}$=10.73, p<0.00001). If cases with regressive ASD do exhibit less delayed early development, but have more severe presentation later in life, then

ADI-R evaluations performed in older individuals should indicate lower severity scores on the ‘abnormality of development evident ≤ 36 months’ domain score, but greater severity scores on other ADI-R domains. Our results connecting immune system function uniquely with phenotypically-defined ASD subgroups support the idea that immune dysfunction is not linked with all forms of ASD, but is confined to specific subphenotypes of ASD[37].

Our results indicate applying a pathway-based approach to analysis of genome-wide ASD data helps account for underlying genetic heterogeneity. This was apparent when comparing genes in the same biological pathway that were associated with subgroups from the two independent datasets. For example, results from the AGRE analyses show the ‘taste transduction’ pathway is very significant ($p<0.001$) when case status is defined using solely diagnostic criteria (Pathway Analysis 1), but not when more extensive phenotype definition is used to classify ASD subgroups ($p_{LessSevere}=0.018$; $p_{MoreSevere}=0.124$). Upon further investigation of core genes driving the association with this pathway, we see unique genomic features associate ($p<0.001$) with the ‘less severe’ subgroup (*ADCY4*, *PRKACA*, *TAS2R16*, *ADCY8*) and others the ‘more severe’ (*KCNB1, GNAS, TAS2R13, TAS2R14, TAS2R43, TAS2R31, TAS2R46, TAS2R19, TAS2R20, TAS2R50, TAS2R42*). While the same exact genes have not to our knowledge been previously implicated in ASD, the chromosomal locations coding these genes have been found linked to male-only subgrouped phenotypes[205] and affected sib-pairs[97]. Also, a SNP near the *TAS2R1* gene on chr5p15 was identified in a GWAS of exclusively multiplex families[221]. Neither the specific taste receptor gene nor assigned SNPs were significant ($p_{TAS2R1}=1.000$; $p_{SNPs}\geq0.0595$) in our studies. One set of taste-related genes appear to be working in the 'less severe' subgroup, and another subset in the 'more severe' subgroup. It is conceivable that multiple different genes functioning in one, or a few pathways, could lead to many different phenotypic consequences, culminating in the autistic spectrum. There is substantial evidence supporting this concept in ASD, reviewed in Geshwind, 2008[69].

Another example of how genetic heterogeneity was accounted for is the association of the pathway described as ‘Bacterial invasion of epithelial cells’. This pathway is very significant ($p \leq 0.002$) when case status is defined using solely diagnostic criteria for both datasets. When affected individuals from both datasets are further defined into ‘less severe’ subgroups, this association signal is no longer significant ($p_{\text{AGRE‘LessSevere’}}=0.538$, $p_{\text{AGP‘LessSevere’}}=0.689$). However, when affected individuals from the two datasets are further defined into ‘more severe’ subgroups the pathway association remains significant ($p_{\text{AGRE‘MoreSevere’}}<0.001$, $p_{\text{AGP‘MoreSevere’}}=0.021$). Upon further investigation of 'core' genes in this pathway, we observed that different genes were associated with the 'more severe' AGRE subgroup when compared to the 'more severe' AGP subgroup. While the same specific gene did not validate, genes within the same family and different genes with similar predicted function related to single transduction and cell motility were identified as driving the pathway's association with both datasets. For example, the *ARPC3* and *ARPC5* genes were significantly associated with the ‘more severe’ AGRE subgroup while the *ARPC1A* gene was significantly associated with the ‘more severe’ AGP subgroup. In a typical single-SNP approach to analysis of GWAS data, or candidate gene analyses, the validated association of this mechanism with the AGP dataset would have gone unnoticed. Known functions of core genes driving the associations for this pathway in the ‘more severe’ subgroups relate to single transduction and cell motility, processes crucial to proper neurodevelopment. Interestingly, some of these core genes have been previously linked to ASD, and in some cases with specific endophenotypes. For example, the genomic region encoding *ARPC5L* (9q33-q34) was found linked in multiplex families when using ‘age at first word’ from the ADI-R as a quantitative trait[179]. This item is included in calculating the ‘abnormality of development evident ≤ 36 months’ domain score and a majority of cases with very severe scores for this measure are assigned to the ‘more severe’ ASD subgroups.

Other interesting pathways identified initially when analyzing all diagnostically-defined cases, upon subgrouping, appear to be uniquely associated with the ‘more severe’ AGRE subgroup ($p_{\text{’MoreSevere’}}<0.001$; $p_{\text{’LessSevere’}}\geq 0.538$). The ‘Neurotrophin signaling pathway’ validated, however exclusively in the ‘less severe’ AGP subgroup ($p_{\text{’LessSevere’}}=0.003$; $p_{\text{All}}=0.168$; $p_{\text{’MoreSevere’}}=0.106$). Many of the core genes driving the pathway association in the ‘more severe’ AGRE subgroup have previously been implicated in nonverbal ASD subgroups. Linkage at interval chr1p13–q12 to nonverbal cases was originally observed in multiplex AGRE families[41]. We identified three core genes in the ‘Neurotrophin signaling pathway’ located within this interval (*JUN*, *NRAS*, and *NGF*). The nerve growth factor (*NGF*) gene is also a core gene in the ‘Apoptosis’ pathway which was uniquely-associated (p<0.001) with the ‘more severe’ AGRE subgroup. Fine-mapping in the previously linked chr1p13-q12 interval detected associations for three haplotype blocks, intronic to the *NGF* gene, in more AGRE families[131]. Further studies identified an association to the *NGF* gene region in an AGP dataset, having simplex and multiplex families. However, this association was to a different haplotype block than the associated AGRE haplotypes and LD calculations indicated these signals were independent. We did not validate direct association to the *NGF* gene in the AGP dataset evaluated in our analyses ($p_{\text{All}}=0.134$; $p_{\text{’LessSevere’}}=0.612$; $p_{\text{’MoreSevere}}$’$=0.145$). Interestingly, there are proportionally more nonverbal individuals in the AGRE subgroups compared to the AGP subgroups ($z_{\text{’LessSevere’}}=5.09$, p<0.00001; $z_{\text{’MoreSevere’}}=15.88$, p<0.00001). These results may further support a relationship between variations in the *NGF* gene and nonverbal ASD subgroups.

While we were able to validate a portion of pathways and core genes identified in the AGRE dataset in the AGP dataset, numerous pathways either do not validate, or validate in a different subgroup classification. Despite our success developing and applying novel multivariate statistical methods to identify genetically meaningful ASD subgroups in both datasets, there are still substantial phenotypic differences between the AGRE and AGP datasets and similarly-defined subgroups from the two datasets. These differences are potentially why we observe distinct genetic signals when comparing results for similarly-defined subgroups from the two datasets. For instance, recent research has suggested that both the phenotypic expression and underlying genetic architecture of ASD in multiplex families is distinct from that in simplex families[216]. Many of the previously reported candidate genes we found associated with the 'more severe' AGRE subgroup, that do not validate in the 'more severe' AGP subgroup, were initially identified in exclusively multiplex families, or in analysis of subgroups from the AGRE dataset. The majority of families in the evaluated AGRE dataset are multiplex (91.2%) compared to a minority of families in the evaluated AGP dataset (31%). Kruskal-Wallis tests show family structure is significantly different ($p<0.0001$) when comparing both the 'less severe' subgroups and 'more severe' subgroups defined in both datasets. The proportion of multiplex families evaluated in AGRE subgroups was also significantly higher than in AGP subgroups ($z_{\text{'LessSevere'}}=15.28$, $p<0.00001$; $z_{\text{'MoreSevere'}}=15.35$, $p<0.00001$). There is also previous evidence indicating sex-specific genetic effects underlying ASD[129]. Similar to our observations regarding dataset-specific family structure, gender is very different between the AGRE 'less severe' subgroup and the AGP 'less severe' subgroup ($p=0.0063$) as well as the AGRE 'more severe' subgroup and the AGP 'more severe' subgroup ($p=0.0021$). The proportion of females evaluated in AGRE subgroups was significantly higher than in AGP subgroups ($z_{\text{'LessSevere'}}=2.73$, $p=0.003$; $z_{\text{'MoreSevere'}}=2.15$, $p=0.016$). It is also notable that the AGRE and AGP samples were genotyped on two different microarray SNP platforms. 3.4% of the pathway-analyzed SNPs identified in the AGRE dataset were not genotyped in the AGP dataset.

We observed strong associations for nine sex chromosome SNPs in the AGP single-SNP analyses. Five SNPs located on the pseudoautosomal region of the X/Y chromosomes, and four SNPs located on the X chromosome pass the threshold for Bonferroni-adjusted significance ($p \leq 6.42x10^{-8}$) (Figure 3.8; Table 3.6). It is difficult to assess the validity of these very significant SNPs. This is mainly due to the statistical limitations involved in evaluating associations for sex-specific genetic markers. These markers were also not assayed on the platforms used to genotype the discovery AGRE dataset. We performed pathway-based analyses with and without these SNPs included and saw no appreciable effects on the significance of associations. This is not unexpected, two of these SNPs are not within +/-50kb of any predicted gene boundaries, one SNP is assigned to a pseudogene (*SSX6*), and one SNP is assigned to a long intergenic non-protein coding RNA (XR_110926.1). For the remaining SNPs, there is previous evidence supporting the involvement of the assigned genes in underlying mechanisms of ASD. SNP rs2896799 ($OR_{All\_AGP} \approx 6.19$; 95% CI=3.77-10.16) is located inside gene boundaries for both *KAL1* and *VCX3B*. *KAL1* is predicted to be involved in neurite outgrowth, axon guidance and branching, and cell adhesion[55]. All developmental mechanisms thought to be involved in ASD[96, 152, 202]. The involvement of *VCX3B* in ASD etiology has also been implicated via inherited deletions of this genomic region[43]. SNP rs909439 ($OR_{All\_AGP} \approx 12.72$; 95% CI=5.54-29.20) is located in *VAMP*7, a gene also known to be involved in neurite outgrowth[9, 10]. Two SNPs, rs34013457 ($OR_{All\_AGP} \approx 5.24$; 95% CI=3.21-8.56) and rs34537684 ($OR_{All\_AGP} \approx 7.53$; 95% CI=4.47-12.69), are located in *PCDH11*. This gene is a member of the protocadherins family. Other genes in the protocadherins family have been previously implicated statistically, and via their functions in synaptic cell-adhesion pathways[27].

The pathway-based approach seems to be a more biologically relevant way to evaluate the effects of common, single genetic variants, especially in a group of disorders known to be as complex and heterogeneous as ASD. We show our method of phenotypic subgrouping is genetically relevant and that using a pathway-based approach to evaluate genetic effects on ASD risk is an effective way to account for genetic heterogeneity, implicating more refined biological mechanisms. By further linking functional pathways of interest and refining the genetic bottlenecks effecting proper pathway function related to specific ASD traits, there may be potential to discover more effective methods of symptom treatment.

CHAPTER IV

EVALUATING SMALL MOLECULE EFFECTS ON EXPRESSION OF AN AUTISM CANDIDATE GENE: *ACETYLSEROTONIN O-METHYLTRANSFERASE*

Introduction

Uncovering pathways associated with subgroups of ASD has elucidated potential sets of genes involved in expression of certain ASD traits. However, to progress toward understanding how these significant findings contribute to disorder process, further functional characterization of these associations is necessary. Most genes identified through pathway analysis have some known biological function but the relationship of these genes to ASD is likely unknown. While some progress has been made, there is still much to learn about pathophysiology and pharmacology in ASD[146, 217].

Many children with ASD are currently treated with medical interventions, yet little evidence exists to support the benefit of these treatments[146]. Evidence also supports significant exhibition of adverse side effects of many medications thereby limiting their use to certain ASD patients[78, 110, 146, 185]. The emerging field of pharmacogenetics is concerned with studying the effects of genetic factors on drug response. Previous pharmacogenetic studies suggest that the altered efficacy and varied side effects seen with many drugs used to treat neurological disorders are related to individual genetic variation[109, 124, 174]. Furthermore, evidence from a study evaluating antidepressant efficacy suggests that single nucleotide polymorphisms (SNPs) located in promoter regions directly affect patient response to drug treatment[120]. Single base-pair changes in the genetic code could allow or disrupt binding of small molecule compounds, causing a drug response in a patient with this variant different from that observed in individuals without these changes. Screening for gene expression effects of small molecule compounds has been used previously toward compound profiling and lead discovery[144, 211]. We applied this concept to functional characterization of known ASD-related SNPs to determine if they cause the gene to respond differently to small molecule compounds when compared to the genotype not associated with ASD.

Acetylserotonin O-methyltransferase (*ASMT*), also known as Hydroxyindole-O-methyltransferase, is the initial candidate gene we chose to test for genotype-specific altered gene expression effects *in vitro* when cell lines are exposed to small molecule compounds. *ASMT* encodes the enzyme that catalyzes the final reaction in melatonin synthesis. Numerous studies have reported abnormal levels of melatonin in individuals with ASD[175] and sleep disorders are common in patients with the disorders with prevalence estimates ranging from 39-80%[73, 111, 137, 190]. Melatonin is involved in regulating the sleep-wake cycle in humans and is synthesized in the pineal gland[5, 33, 140]. Synthesis of melatonin begins with the active uptake of the amino acid tryptophan into the gland. Tryptophan is then hydroxylated and decarboxylated to serotonin, another molecule with ample evidence for involvement in ASD[83]. Serotonin is then *N*-acetylated by the rate-limiting enzyme in this pathway, arylalkylamine, and subsequently converted to melatonin by the ASMT enzyme[5].

Melatonin supplementation is an emerging approach to treating sleep defects in ASD, however some patients are non-responders[175]. Other patients undergoing melatonin treatment report relief from comorbid symptoms like irritable bowel syndrome[220], anxiety and seizures[189], while some exhibit more severe symptoms[23, 78, 185]. These seemingly contradictory findings suggest that underlying genetic architecture may affect exhibition of adverse side effects resulting from melatonin treatment. These findings are not exclusive to treatment with melatonin. Interestingly, for many other compounds used to treat comorbid symptoms of ASD, individuals report sleep problems as adverse side effects[146]. One possible explanation is that these small molecule compounds are somehow perturbing the melatonin synthesis pathway, potentially by affecting expression of *ASMT*.

The involvement of *ASMT* in ASD etiology has been studied extensively[99, 148, 175, 219]. There are three isoforms of the *ASMT* gene resulting from alternative splicing of exons 6 and 7[51]. There are also two distinct putative promoters reported, promoter A and promoter B[171]. Previous tissue-specific expression studies indicate promoter A is expressed almost exclusively in the retina, while promoter B drives *ASMT* expression in high amounts in the pineal gland. There are two SNPs located in promoter B of *ASMT* that have been statistically associated with increased ASD risk, rs4446909 and rs5989681. Additionally, homozygous presence of the risk alleles for both SNPs was correlated with a significant decrease in *ASMT* expression and *ASMT* enzymatic activity in patients[148]. The *ASMT* promoter polymorphisms conferring risk for ASD are located in transcription factor binding sites for nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) and specificity protein 1 (Sp1)[99]. As such, the reported SNPs are thought to alter gene expression by disrupting transcription factor binding. An ASD-risk haplotype has also been reported that includes the promoter B SNPs and a third SNP, rs6644635, located in the 5'-untranslated region (UTR) of the only know functional isoform of ASMT[30, 148, 219].

We hypothesized that the *ASMT* gene promoter B could be a target for small molecule compounds and wanted to determine the effects of current ASD treatments on genotype-specific *ASMT* expression. The goal is to determine if effects of individual genetic variation, in relation to *ASMT* expression, could help explain the observed inefficacy and adverse side effects of certain drugs used to treat ASD comorbid symptoms. The ultimate goal for all pharmacogenetic studies is to provide evidence useful toward optimizing more effective medical treatments for each person’s unique genetic architecture.

Methods

Choice of Cell Type

We used previously generated lymphoblast cell lines derived from individuals ascertained by our lab and collaborators. We chose to utilize lymphoblast cell lines (LCLs) to allow evaluation of small molecule effects in the endogenous melatonin system. Promoter B is reported to be actively expressed in LCL and these cells are the same lines used by Melke et al, 2008 to identify the published *ASMT* genotype-specific gene expression, indicating gene expression of the candidate gene should be detectable in these cells. Further, melatonin biosynthesis has been reported in human mononuclear lymphocytes[59]. It is also important to note, LCLs have a relatively low reported somatic mutation rate at low passages (0.3%)[186].

Sequence Confirmation

We screened DNA previously extracted from the blood of 22 individuals, in 15 ASD families, previously genotyped at the rs4446909 marker, for which cell lines were available. A region of *ASMT*, including the promoter B element, 5-UTR, and exon 1B, was amplified for each DNA sample via polymerase chain reactions (PCR) using the following primers: forward 5'-AAAAGGGGTCTCACTATGTTGC-3'; reverse 5'-TGGAACGTGAGTGTGATG AAC-3'. Amplified products were purified from reactions with the QIAquick® PCR Purification Kit and Sanger sequenced at GenHunter® Corporation. Presence of the genotypes of interest at each SNP in the haplotype of interest was verified by analyzing raw sequence chromatograms. The linkage disequilibrium (LD) map for SNPs of interest in this region was calculated using pairwise D' with Haploview.

Cell Culture, DNA and RNA Isolation

We chose two cell lines from affected individuals homozygous for the associated risk haplotype (rs4446909$_{GG}$, rs5989681$_{GG}$, rs6644635$_{CC}$), and one cell line from an affected individual homozygous for the promoter B risk alleles and heterozygous at the third SNP in the haplotype (rs4446909$_{GG}$, rs5989681$_{GG}$, rs6644635$_{CT}$). Two cell lines were also chosen from affected individuals heterozygous at all SNPs (rs4446909$_{AG}$, rs5989681$_{CG}$, rs6644635$_{CT}$), and one cell line from an affected individual heterozygous for the promoter B risk alleles and homozygous at the third SNP in the haplotype (rs4446909$_{AG}$, rs5989681$_{CG}$, rs6644635$_{CC}$). Finally, we chose three cell lines from individuals, two affected and one father, homozygous for the unassociated promoter B genotypes (rs4446909$_{AA}$, rs5989681$_{CC}$, rs6644635$_{CC}$). Due to the lower frequency of these genotypes in our case population, it was necessary to choose one parental cell line. It was previously reported that individuals with homozygous non-risk genotypes at the promoter B SNPs had higher *ASMT* transcription regardless of case status. It was also shown that in parents of children diagnosed with ASD, *ASMT* transcription correlated with melatonin levels[148].

Cells were grown at 37°C in RPMI-1640 medium, plus L-glutamine (Life Technologies, Inc., Grand Island, NY, USA). Growth media was supplemented with 10% heat-inactivated, undialyzed fetal bovine serum (FBS), and 1% penicillin/streptomycin (10,000ug/ml) antibiotic. DNA was extracted using the DNeasy® Tissue Kit from Qiagen®. DNA extracted from cell lines was sequenced, as described above, to verify the correct sequence of interest in each line. Total RNA was isolated using the phenol/chloroform method.

Characterization of Basal ASMT Transcript Levels

Over the course of 4 weeks, at one week intervals, RNA was extracted from each cell line. Oligo(dT)-primed cDNA was constructed from 5µg total RNA, using the Superscript II kit (Invitrogen, Grand Island, NY, USA), according to the manufacturer instructions, with RNase inhibitor. These cDNAs were standardized to the same concentration (100ng) and used directly in quantitative real-time PCR (qPCR). Multiplex qPCRs were performed, in triplicate, using the TaqMan® Fast Advanced Master Mix, on the Applied Biosystems® 7900HT Fast Real-Time PCR System. *ASMT* mRNA was quantified using a commercially available FAM-labeled TaqMan® assay spanning the boundary between exon 1B and exon 2 (Hs00946625_m1). Relative quantification of *ASMT* expression was determined using the comparative cycle threshold ($2^{-\Delta\Delta Ct}$) method. Amplification efficiencies were determined using linear regression analysis performed on log fluorescence data (i.e. the inverse log of the slope in the log linear phase)[66]. *ASMT* expression was normalized to C_t values for a VIC-labeled TaqMan® assay spanning exons 1 and 2 of the polymerase (RNA) II (DNA directed) polypeptide A (*POLR2A*) gene (Hs00172187_m1). Statistical significance of qPCR results was determined using a Student's two-tailed *t*-test, with unequal variance.

Effects of FBS Serotonin Exposure on ASMT Expression

We controlled for serotonin present in the FBS used for cell culture by adapting the cells into completely serum-free media and serum-starving them for at least 24 hours prior to performing small molecule treatments. We used AIM V® Medium, Liquid with Human Serum Albumin. To determine the effect of serum starvation on expression of *ASMT*, aliquots of cells from each line were spun down and resuspended in either the serum-supplemented 'growth' media or the serum-free 'starvation' media, to mimic the experimental environment of small molecule treatments. Resuspended cells were allowed to grow for 24hrs and total RNA was isolated. Oligo(dT)-primed cDNA was constructed and RT-qPCRs were performed as described above.

Small Molecule Treatments

We focused small molecule experiments on five compounds currently used to treat symptoms in ASD, where reported side effects include sleep disturbances. These compounds were: Risperidone, Escitalopram, Fluoxetine, Serotonin and Melatonin. Prior to cell treatments, we ensured receptors for chosen compounds were expressed in human lymphocytes[125, 139, 169, 191]. Cell lines were spun down and resuspended in serum-free media as described above. After at least 24 hours of serum-deprivation, when cells were in the mid-logarithmic phase of growth, six wells were plated for each cell line, at 2.5 ml total volume per well. Small molecule treatments were performed with cells suspended in serum-free media. Experiments were standardized to have similar counts of cells/mL in each treatment well (i.e. cell counts were diluted to equal the well with the lowest cell count/mL and were ~500,000). Compounds were dissolved in DMSO+H_2O and added to cells at concentrations comparable to clinical dosage, when available. For FDA approved drugs, treatment concentrations were determined based on reported peak plasma concentrations in humans[11, 167, 226]. Melatonin has yet to be approved by the FDA, however, pharmacokinetics of melatonin have been reported in older adults[74] and a phase I trial has been performed to evaluate melatonin treatment for sleep problems in autistic individuals[137]. We used the reported effective dosages in autistic children at the corresponding reported peak plasma concentration from the study performed on older adults. For serotonin treatments, we used mean whole-blood 5-HT concentrations reported for children with ASD, which were shown to be higher when compared to healthy control children[80]. Negative controls were treated with vehicle-only (DMSO+H_2O) (Table 4.1). Six hours after addition of compounds, 1mL of cells from each treatment were aliquoted, spun down at 1,000g for 5 minutes, and frozen at -80°C. The remaining 1.5mL were spun down at 1,000g for 5 minutes, and frozen at -80°C, 12 hours after addition of compounds. This treatment protocol was performed in three experimental replicates over the course of one week to minimize potential biases that may arise due to different batches of growth media and serum-free media, and cell passages. Total RNA was isolated from cells from the first small molecule treatment experiments. Randomly-primed cDNA was constructed, without RNase inhibitor, using the High-Capacity cDNA Reverse Transcription Kits from Applied Biosystems. RT-qPCRs were performed, in triplicate, as described above.

Compound	Drug (g)	DMSO (ml)	Compound Initial Concentration (ng/ml)	Final Concentration (ng/ml)	Final Volume Added to Cells (ul)
Risperidone	0.0005	0.1	50000	15.90	0.80
Melatonin	0.0001	1.0	150000	18.80	0.31
Fluoxetine	0.0006	0.1	62500	171.00	6.84
Escitalopram	0.0015	0.1	170000	278.80	4.10
Serotonin	0.0017	0.1	1000	4.00	10.00
Negative Control: DMSO+H_2O	NA	1.0	NA	1ml DMSO:100ml H2O	10.00

Table 4.1. Compound Dilutions for Cell Line Treatments. Reported are the final concentrations and amounts of small molecule compounds added to cell lines for experimental treatments. All compounds were initially dissolved in the recommended amount of DMSO. Initial compound dilutions were then diluted further with 100mL H_2O to allow pipettable volumes for experiments.

Results

Sequence Confirmation

Sequencing of the *ASMT* promoter B element and 5'-UTR for the 21 affected individuals (15 males and 7 females), we evaluated indicates low levels of LD across the three markers previously reported to be inherited as a risk haplotype for ASD (Figure 4.1).

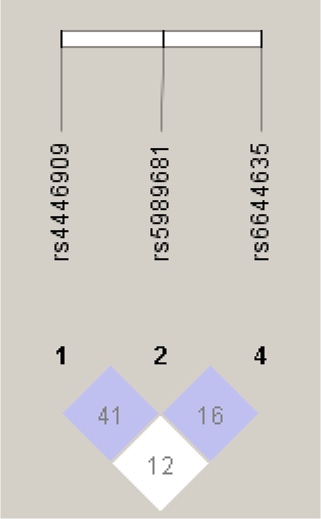


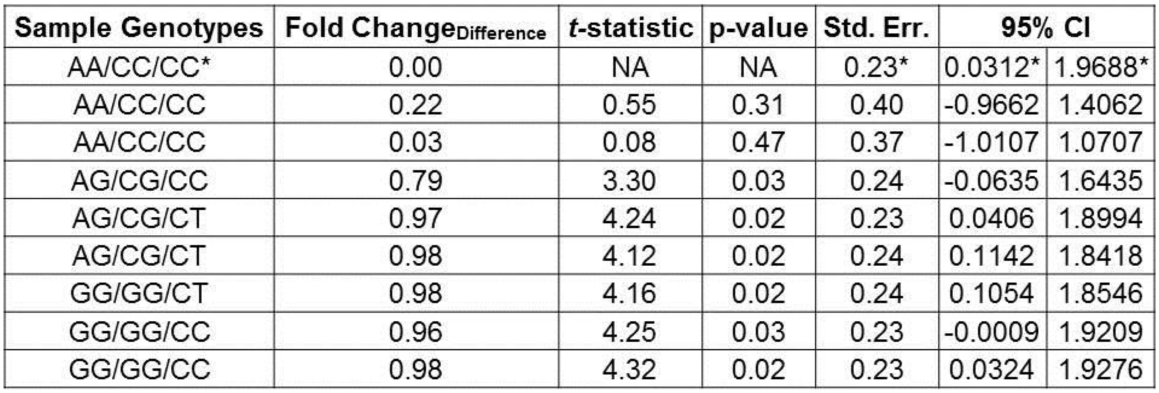
Figure 4.1. Haplotype block structure of the promoter B and 5'-UTR SNPs in *ASMT*. Reported are the relative position of each SNP, and the pairwise LD (r^2) between all SNPs.

Characterization of Basal ASMT Transcript Levels & Effects of FBS Serotonin Exposure on ASMT Expression

Results from qPCR for these experiments show C_ts vary widely across triplicates and reactions have low amplification efficiencies (efficiencies < 80%). Evaluation of the raw amplification plots show that expression of the endogenous control gene we chose, *POLR2A*, is extremely variable across triplicates, and in many cases has lower C_t-values than the gene of interest.

Small Molecule Treatments

Amplification efficiencies for qPCR are low for a majority of these reactions. Efficiencies for negative controls treated with vehicle-only range from 66%-128% for the *ASMT* assay, and 73%-83% for the *POLR2A* assay. We potentially see significantly (p≤0.03) decreased *ASMT* expression for individuals homozygous for risk alleles at the promoter B SNPs, compared to individuals homozygous for the alternative alleles at these SNPs. We also observe significant reductions in *ASMT* expression for individuals heterozygous at the two promoter B SNPs, compared to individuals homozygous for the alternative alleles at these SNPs (Table 4.2; Figure 4.2).



Sample Genotypes	Fold Change$_{\text{Difference}}$	*t*-statistic	p-value	Std. Err.	95% CI	
AA/CC/CC*	0.00	NA	NA	0.23*	0.0312*	1.9688*
AA/CC/CC	0.22	0.55	0.31	0.40	-0.9662	1.4062
AA/CC/CC	0.03	0.08	0.47	0.37	-1.0107	1.0707
AG/CG/CC	0.79	3.30	0.03	0.24	-0.0635	1.6435
AG/CG/CT	0.97	4.24	0.02	0.23	0.0406	1.8994
AG/CG/CT	0.98	4.12	0.02	0.24	0.1142	1.8418
GG/GG/CT	0.98	4.16	0.02	0.24	0.1054	1.8546
GG/GG/CC	0.96	4.25	0.03	0.23	-0.0009	1.9209
GG/GG/CC	0.98	4.32	0.02	0.23	0.0324	1.9276

Table 4.2. Fold Change Differences By Genotypes. Reported are results from Student's *t*-tests, with unequal variance. Genotypes are indicated in order of chromosomal location: rs4446909/rs5989681/rs6644635. Asterisks denote calibrator sample, statistics for this sample are reported for the mean calculated across triplicates. All other statistics represent those calculated for the fold change difference observed. Std. Err.=standard error, 95% CI=95% confidence interval.

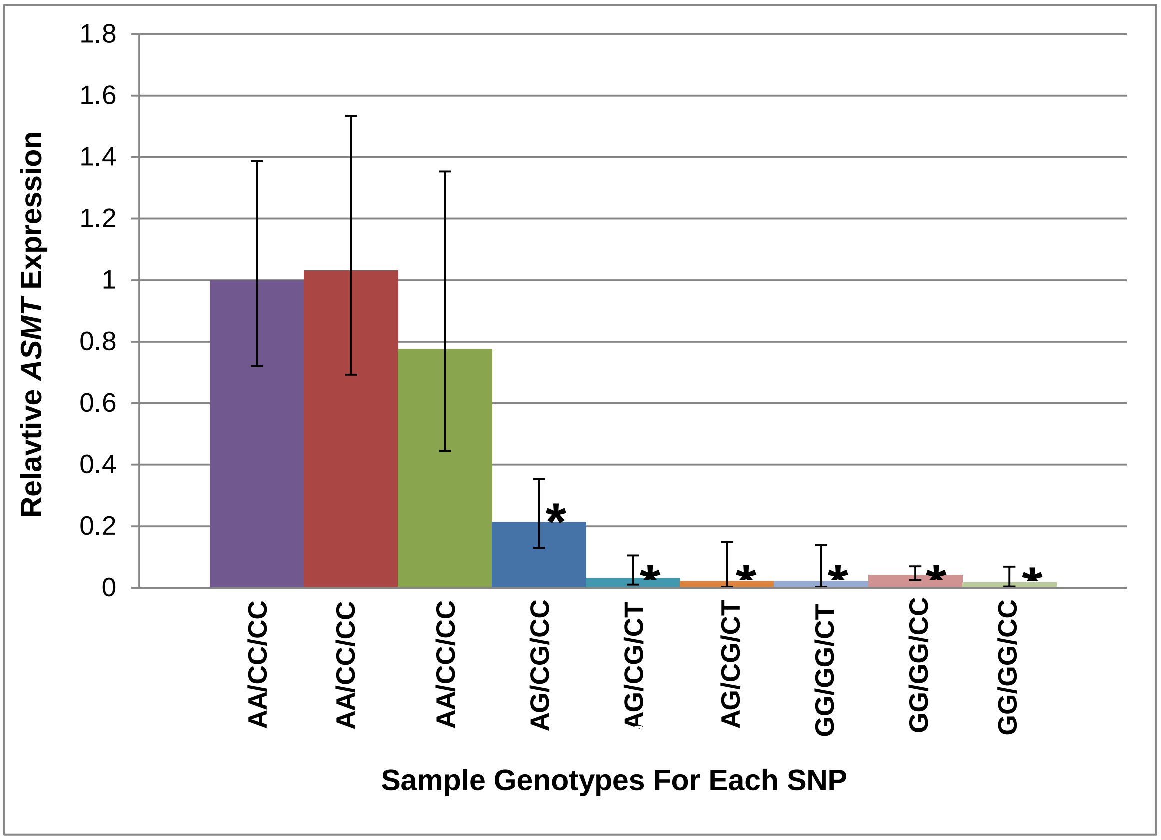


Figure 4.2. Genotype-Specific *ASMT* Gene Expression. Quantification of *ASMT* transcripts relative to genotypes of interest for each SNP. Genotypes are indicated in order of chromosomal location: rs4446909/rs5989681/rs6644635. Statistical significance determined via Student's *t* test with unequal variance. *****p ≤ 0.03.

We observed one sample having a potentially significant increase (p=0.02) in *ASMT* expression after exposure to Serotonin for 12 hours. This did not replicate across the other two samples with the same combination of genotypes at the three SNPs of interest (AA/CC/CC). We did not observe any other significant changes in *ASMT* expression following exposure of cells with the non-risk genotypes to any of the other evaluated compounds (Figure 4.3). Results from qPCR for treatment experiments on samples heterozygous or homozygous for risk genotypes vary widely across triplicates and are inconclusive.

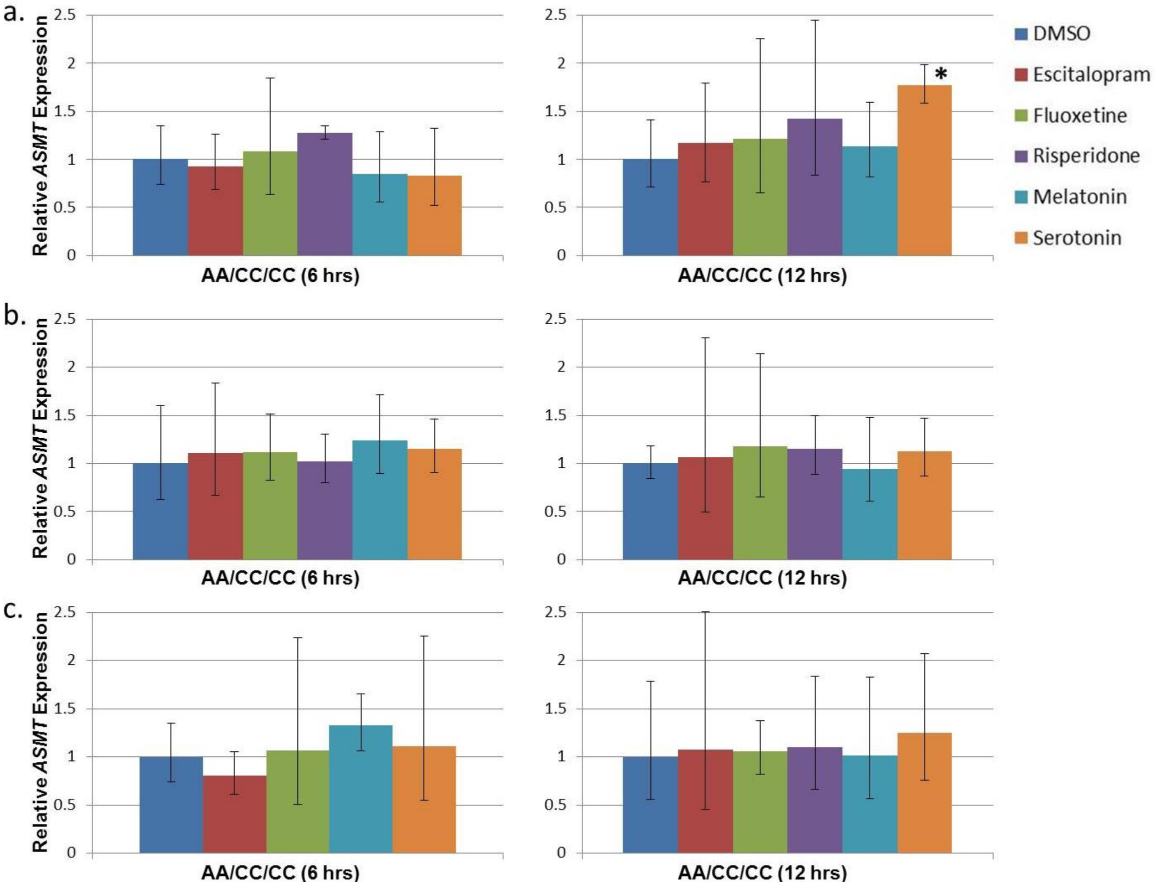


Figure 4.3. Effects of Small Molecule Treatments on *ASMT* Gene Expression: Non-Risk Alleles. Quantification of *ASMT* transcripts following exposure to evaluated small molecule compounds for the non-risk haplotypes. Genotypes are indicated in order of chromosomal location: rs4446909/rs5989681/rs6644635. Statistical significance determined via Student's *t* test with unequal variance. *****p = 0.03.

Discussion

We observe higher recombination rates between SNPs in the predicted risk haplotype for our small subset of samples, compared to currently reported estimates in larger European and Han Chinese descent ASD cohorts[148, 219]. The previously reported structure for the *ASMT* promoter B SNPs and 5'-UTR SNP suggests strong linkage disequilibrium (LD) across all three markers. D' estimates between the two promoter B SNPs, rs4446909 and rs5989681, suggest the two SNPs are inherited more often as a haplotype (D’=0.92-0.94). This is also true for reported D' estimates between rs4446909 and rs6644635 (D'=0.84), and rs5989681 and rs6644635 (D'=0.98)[148, 219]. We are reporting pairwise measures of the squared correlation coefficient (r^2), since D' calculations in our evaluated samples are not informative. Due to the equation used to calculate the D' statistic, missing genotype combinations always result in D'=1. There are missing genotype combinations between markers rs4446909 and rs598968, and markers rs598968 and rs6644635. Our estimates of r^2 suggest these three SNPs are not inherited as a haplotype block in these ASD families.

Unfortunately, we were unable to ensure that the level of *ASMT* expression in our cell lines was stable over a four-week time course. However, previous evidence suggests *ASMT* mRNA expression and enzymatic activity in the pineal gland does not fluctuate based on diurnal rhythms[4]. We were also unable to determine the effect of serum-deprivation on expression of *ASMT*. Interestingly, the observation that *POLR2A* had lower expression than *ASMT* for most samples was only in reactions using the Oligo(dT)-primed cDNA prepared with an RNase Inhibitor. There is previous evidence suggesting decreases in *POLR2A* mRNA levels are attributable to RNase H-mediated cleavage of the mRNA[206]. It is possible that by treating cDNA samples with RNase H we affected expression of our endogenous control, rendering the qPCR results unreliable. We evaluated seven different potential control genes prior to performing these qPCR experiments. The goal was to obtain a normalizer gene with C_t-values similar to those observed for *ASMT*. An alternative gene we anticipate using in future experiments is *GAPDH*.

By evaluating qPCR results from vehicle-only treated controls, our data are consistent with previous findings indicating homozygous presence of the risk alleles at the promoter B SNPs, rs4446909 and rs5989681, results in decreased *ASMT* gene expression[148]. Previous reports also suggest the observed decrease in expression was only attributable to homozygosity for risk alleles at these two SNPs. There are no reported effects on *ASMT* gene expression attributable to heterozygosity at the promoter B SNPs or to any genotype at the 5'-UTR SNP, rs6644635. We do observe decreased *ASMT* expression when individuals are heterozygous, compared to homozygous non-risk genotypes at these markers. This is very preliminary and to accurately determine the effects of heterozygosity at these markers on *ASMT* expression would require further experiments aimed at modeling the effect of genotypes at each SNP alone and conducted on cDNA extracted from entirely untreated cells.

Initial results suggest there are no large changes in *ASMT* gene expression upon exposure to small molecule compounds at either the 6 or 12 hour time point for the non-risk haplotype. Again, reaction efficiencies are low and estimates of relative *ASMT* quantities are highly variable across triplicates. This is especially true for samples where *ASMT* transcript production is already reduced in negative controls. It is possible the low level of expression for our candidate gene in LCL is too low to be accurately detected via qPCR. It is also possible that exposure to the small molecule compounds alter expression of our chosen endogenous control gene. The low reaction efficiencies could also be attributable to pipetting error, poor PCR primer design, a result of multiplexing the reaction, or cDNA concentrations that are too low, or high, to be detected accurately[108]. It is difficult to determine the effects of small molecule compound exposure on *ASMT* expression using these reported results. Future experiments, directed at optimizing the qPCR, will be necessary to formulate conclusions from our small molecule treatments.

CHAPTER V

CONCLUSION

Summary

Autism Spectrum Disorder exhibits multiple levels of complexity related to clinical manifestation and etiology. There are many mechanisms implicated in ASD, including, but not limited to, biological epistasis, genetic heterogeneity, gene-environment interactions, and epigenetic effects. The research conducted in this dissertation was motivated by the idea that the difficulty in identifying genetic variation with strong effects on risk for ASD is due to the wide variability in clinical manifestation, being explained in large part by underlying genetic heterogeneity.

We hypothesized that phenotypic heterogeneity could be one phenomenon complicating identification of genetic factors. By performing unsupervised clustering, based on a myriad of carefully chosen phenotypic information, derived from more than one source, we were able to effectively evaluate a broad array of information and enable a more complete phenotype definition for subsets of individuals with ASD. The overlapping interpretation of our results from two different multivariate analyses, PCA and clustering, demonstrate the utility of this approach. That we were able to show defined subgroups of phenotypic expression appearing to be genetically meaningful in the AGRE dataset and replicate these findings in an independent AGP dataset lends further support to the validity of the resulting cluster groupings and the idea that the phenotype clusters recapitulate underlying genetic mechanisms in Autism Spectrum Disorders.

To further support this idea, we see that unique biological mechanisms are implicated when comparing genes associated with either the ‘more severe’ or ‘less severe' ASD subgroups. Our results suggest that meaningful phenotypic subgroup definitions can help clarify the underlying genetic etiology of Autism Spectrum Disorders. The pathway-based approach seems to be a more biologically relevant way to evaluate the effects of common, single genetic variants, especially in a group of disorders known to be as complex and genetically heterogeneous as ASD. We show that using a pathway-based approach to evaluate genetic effects on ASD risk is an effective way to account for genetic heterogeneity, implicating more refined biological mechanisms. By further linking functional pathways of interest and refining the genetic bottlenecks effecting proper pathway function related to specific ASD traits, there may be potential to discover more effective methods of symptom treatment.

Results from our functional pharmacogenetic analyses evaluating genotype-specific small molecule effects on expression of *ASMT* are largely inconclusive and will need to be evaluated further in future studies. However, our data are consistent with previous results indicating homozygous presence of risk alleles at the promoter B SNPs significantly reduces *ASMT* expression. We also have potentially implicated previously unreported gene expression effects related to heterozygosity. We did not observe any conclusive effects of compound treatment on expression in the non-risk haplotype. This may indicate that altered efficacy and presentation of adverse sleep-related events are not attributable to deregulation of *ASMT*.

Future Directions

We chose to conduct completely separate analyses in the AGRE and AGP datasets. The initial goal was to determine the replicability of the phenotypic subgrouping. As such, it was necessary to run independent multivariate statistical analyses in these datasets. However, the utility of this approach may not have been the most powerful option in our subsequent genetic analyses. There are many potential reasons that a portion of the sub-group specific genetic results, identified in the AGRE dataset, did not validate, or validate in a different subgroup, in the AGP dataset. A number of these potential reasons are discussed in more detail in Chapter III. To truly replicate genetic analyses in the AGP dataset, it will be necessary to perform a confirmatory factor analysis by applying the same cluster analysis and principal component loadings identified in the AGRE dataset, to the AGP dataset. In other words, to fit phenotype characteristics of individuals in the AGP dataset into the defined AGRE clusters.

It is interesting to speculate at potential genotype-phenotype relationships resulting from pathway analyses of the main cluster groupings. However, there is still substantial phenotypic heterogeneity in these main subgroups within the same dataset. It would be beneficial to further evaluate genetic contributions to ASD-related phenotypes in the smaller subclusters, as opposed to main clusters. The defined subclusters within each main cluster seem to represent ASD subgroups with more homogeneous phenotypic expression than the main clusters, and could be very informative for these types of evaluations. For example, an interesting analysis would be to evaluate genetic contributions in the 'youngest' subclusters. These subclusters grouped separately from the other subclusters within the ‘more severe’ main clusters for both datasets.

Since the functional focus of the KEGG database is definition of primarily metabolic pathways, it would be interesting to evaluate other pathway databases more potentially relevant to functional mechanisms implicated in ASD. We have evaluated the Gene Ontology database with PARIS and have numerous interesting results from these analyses that could be evaluated in future studies. Preliminary examination of the Gene Ontology results show that some of the core genes identified in the KEGG database analyses overlap with genes in the Gene Ontology database, but many strongly associated genes are unique.

To obtain more conclusive and reliable results from our small molecule experiments, it will be necessary to further troubleshoot qPCR and try to obtain tighter cycle thresholds for triplicates. Obtaining an endogenous control gene that is not affected by exposure to small molecule compounds is an important next step. In the future, we would like to run qPCR normalizing to a primer-limited assay for *GAPDH*, to determine if compounds do have genotype-specific functional effects that do not relate directly to *ASMT* expression. It would be interesting to evaluate potential expression effects of these small molecules on other genes in the melatonin pathway. It may be that the expression of other genes in the melatonin pathway is dependent on endogenous *ASMT* expression, which is altered due to the genotype-specific effects of SNPs in the *ASMT* promoter B and 5-UTR. It would also be relevant to perform unbiased transcriptome profiling using RNA extracted from our treated cells to evaluate potential expression effects on many genes that function in other pathways, in addition to the melatonin pathway.

The importance of determining the relationship of genotype to phenotype in all aspects of genetic analysis of complex disease cannot be overstated. Most of the ASD risk genes identified, especially via pathway-based analysis, have some known biological function, but the relationship of these genes to ASD is largely unknown. The currently known list of ASD risk genes and other genetic abnormalities need to be extensively studied to truly understand the functional consequences of each variation. While progress has been made, there is still much to learn about pathophysiology and pharmacology in ASD.

REFERENCES

1. Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. *MMWR. Surveill. Summ.* **61**, 1-19 (2012).

2. Abdi,H. & Williams,L. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 433-459 (2010).

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

4. Ackermann,K., *et al.* Characterization of human melatonin synthesis using autoptic pineal tissue. *Endocrinology* **147**, 3235-3242 (2006).

5. Ackermann,K. & Stehle,J.H. Melatonin synthesis in the human pineal gland: advantages, implications, and difficulties. *Chronobiol. Int.* **23**, 369-379 (2006).

6. Alarcon,M., *et al.* Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82**, 150-159 (2008).

7. Alarcon,M., *et al.* Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. *Am. J. Hum. Genet.* **82**, 150-159 (2008).

8. Alarcon,M., *et al.* Quantitative genome scan and Ordered-Subsets Analysis of autism endophenotypes support language QTLs. *Mol. Psychiatry* **10**, 747-757 (2005).

9. Alberts,P., *et al.* Cross talk between tetanus neurotoxin-insensitive vesicle-associated membrane protein-mediated transport and L1-mediated adhesion. *Mol. Biol. Cell* **14**, 4207-4220 (2003).

10. Alberts,P., *et al.* Cdc42 and actin control polarized expression of TI-VAMP vesicles to neuronal growth cones and their fusion with the plasma membrane. *Mol. Biol. Cell* **17**, 1194-1203 (2006).

11. Aman,M.G., *et al.* Plasma pharmacokinetic characteristics of risperidone and their relationship to saliva concentrations in children with psychiatric or neurodevelopmental disorders. *Clin. Ther.* **29**, 1476-1486 (2007).

12. American Psychiatric Association *Diagnostic and Statistical Manual of Mental Disorders, 4th edition, Text Revision*(American Psychiatric Association, Washington DC, 2000).

13. American Psychiatric Association. Report of DSM-5 Proposed Criteria for Autism Spectrum Disorder Designed to Provide More Accurate Diagnosis and Treatment. 2013. Arlington, VA, American Psychiatric Association.
Ref Type: Report

14. Amir,R.E., *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185-188 (1999).

15. Anney,R., *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum. Mol. Genet.* **21**, 4781-4792 (2012).

16. Anney,R., *et al.* Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum. Mol. Genet.* **21**, 4781-4792 (2012).

17. Anney,R., *et al.* A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.* **19**, 4072-4082 (2010).

18. Askland,K., Read,C., & Moore,J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.* **125**, 63-79 (2009).

19. Atladottir,H.O., *et al.* Association of family history of autoimmune diseases and autism spectrum disorders. *Pediatrics.* **124**, 687-694 (2009).

20. Aulchenko,Y.S., *et al.* GenABEL: an R library for genome-wide association analysis. *Bioinformatics.* **23**, 1294-1296 (2007).

21. Bailey,A., *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol. Med.* **25**, 63-77 (1995).

22. Bakermans-Kranenburg,M.J. & van,I.J.M. Sniffing around oxytocin: review and meta-analyses of trials in healthy and clinical groups with implications for pharmacotherapy. *Transl. Psychiatry* **3**, e258 (2013).

23. Banach,M., *et al.* Melatonin in experimental seizures and epilepsy. *Pharmacol. Rep.* **63**, 1-11 (2011).

24. Bartlett,C.W., *et al.* Examination of potential overlap in autism and language loci on chromosomes 2, 7, and 13 in two independent samples ascertained for specific language impairment. *Hum. Hered.* **57**, 10-20 (2004).

25. Beaudet,A.L. Autism: highly heritable but not inherited. *Nat. Med.* **13**, 534-536 (2007).

26. Betancur,C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380**, 42-77 (2011).

27. Betancur,C., Sakurai,T., & Buxbaum,J.D. The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders. *Trends Neurosci.* **32**, 402-412 (2009).

28. Bolte,S. & Poustka,F. The relation between general cognitive level and adaptive behavior domains in individuals with autism with and without co-morbid mental retardation. *Child Psychiatry Hum. Dev.* **33**, 165-172 (2002).

29. Bolton,P., *et al.* A case-control family history study of autism. *J. Child Psychol. Psychiatry* **35**, 877-900 (1994).

30. Botros,H.G., *et al.* Crystal structure and functional mapping of human ASMT, the last enzyme of the melatonin synthesis pathway. *J. Pineal. Res.*(2012).

31. Brock,G., Pihur,V., Datta,S., & Datta,S. clValid: An R Package for Cluster Validation. Journal of Statistical Software 25[4]. 2008.

Ref Type: Computer Program

32. Bruining,H., *et al.* Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. *PLoS. One.* **5**, e10887 (2010).

33. Brzezinski,A. Melatonin in humans. *N. Engl. J. Med.* **336**, 186-195 (1997).

34. Bucan,M., *et al.* Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS. Genet.* **5**, e1000536 (2009).

35. Butler,M.G., *et al.* Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J. Med. Genet.* **42**, 318-321 (2005).

36. Buxbaum,J.D., *et al.* Linkage analysis for autism in a subset families with obsessive-compulsive behaviors: evidence for an autism susceptibility gene on chromosome 1 and further support for susceptibility genes on chromosome 6 and 19. *Mol. Psychiatry* **9**, 144-150 (2004).

37. Careaga,M., Van de Water,J., & Ashwood,P. Immune dysfunction in autism: a pathway to treatment. *Neurotherapeutics.* **7**, 283-292 (2010).

38. Carson,M.J., *et al.* CNS immune privilege: hiding in plain sight. *Immunol. Rev.* **213**, 48-65 (2006).

39. Cartwright,C. & Hollander,E. SSRIs in the treatment of obsessive-compulsive disorder. *Depress. Anxiety.* **8 Suppl 1**, 105-113 (1998).

40. Chen,G.K., *et al.* Quantitative trait locus analysis of nonverbal communication in autism spectrum disorder. *Mol. Psychiatry* **11**, 214-220 (2006).

41. Chen,G.K., *et al.* Quantitative trait locus analysis of nonverbal communication in autism spectrum disorder. *Mol. Psychiatry* **11**, 214-220 (2006).

42. Chisaka,O., Musci,T.S., & Capecchi,M.R. Developmental defects of the ear, cranial nerves and hindbrain resulting from targeted disruption of the mouse homeobox gene Hox-1.6. *Nature.* **355**, 516-520 (1992).

43. Chocholska,S., *et al.* Molecular cytogenetic analysis of a familial interstitial deletion Xp22.2-22.3 with a highly variable phenotype in female carriers. *Am. J. Med. Genet. A.* **140**, 604-610 (2006).

44. Constantino,J.N. & Todd,R.D. Intergenerational transmission of subthreshold autistic traits in the general population. *Biol. Psychiatry* **57**, 655-660 (2005).

45. Cook,E.H., Jr. Genetics of autism. *Child Adolesc. Psychiatr. Clin. N. Am.* **10**, 333-350 (2001).

46. Craig,A.M. & Kang,Y. Neurexin-neuroligin signaling in synapse development. *Curr. Opin. Neurobiol.* **17**, 43-52 (2007).

47. Crespi,B.J. & Crofts,H.J. Association testing of copy number variants in schizophrenia and autism spectrum disorders. *J. Neurodev. Disord.* **4**, 15 (2012).

48. Curatolo,P., *et al.* Autism in tuberous sclerosis. *Eur. J. Paediatr. Neurol.* **8**, 327-332 (2004).

49. Devlin,B. & Roeder,K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).

50. Devlin,B. & Scherer,S.W. Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.* **22**, 229-237 (2012).

51. Donohue,S.J., *et al.* Human hydroxyindole-O-methyltransferase: presence of LINE-1 fragment in a cDNA clone and pineal mRNA. *DNA Cell Biol.* **12**, 715-727 (1993).

52. Dudbridge,F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum. Hered.* **66**, 87-98 (2008).

53. Eaves,L.C., Ho,H.H., & Eaves,D.M. Subtypes of autism by cluster analysis. *J. Autism Dev. Disord.* **24**, 3-22 (1994).

54. El-Fishawy,P. & State MW The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatr. Clin. North Am.* **33**, 83-105 (2010).

55. Engle,E.C. Human Genetic Disorders of Axon Guidance. *Cold Spring Harb Perspect Biol.* **2**, (2010).

56. Evans,D.M. & Purcell,S. Power calculations in genetic studies. *Cold Spring. Harb. Protoc.* **2012**, 664-674 (2012).

57. Fein,D., *et al.* Subtypes of pervasive developmental disorder: Clinical characteristics. *Child Neuropsychology* **5**, 1-23 (1999).

58. Feng,J., *et al.* High frequency of neurexin 1beta signal peptide structural variants in patients with autism. *Neurosci. Lett.* **409**, 10-13 (2006).

59. Finocchiaro,L.M., Nahmod,V.E., & Launay,J.M. Melatonin biosynthesis and metabolism in peripheral blood mononuclear leucocytes. *Biochem. J.* **280 (Pt 3)**, 727-731 (1991).

60. Folstein,S. & Rutter,M. Genetic influences and infantile autism. *Nature.* **265**, 726-728 (1977).

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

62. Fombonne,E., *et al.* Microcephaly and macrocephaly in autism. *J. Autism Dev. Disord.* **29**, 113-119 (1999).

63. Fraley,C. & Raftery,A. Enhanced model-based clustering, density estimation and discriminant analysis software: MCLUST. Journal of Classification 20, 263-286. 2003.

Ref Type: Computer Program

64. Frazier,T.W., *et al.* Validation of proposed DSM-5 criteria for autism spectrum disorder. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 28-40 (2012).

65. Funalot,B., Varenne,O., & Mas,J.L. A call for accurate phenotype definition in the study of complex disorders. *Nat. Genet.* **36**, 3-4 (2004).

66. Gentle,A., Anastasopoulos,F., & McBrien,N.A. High-resolution semi-quantitative real-time PCR without the use of a standard curve. *Biotechniques.* **31**, 502, 504-506, 508 (2001).

67. Georgiades,S., *et al.* Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *J. Child Psychol. Psychiatry*(2012).

68. Geschwind,D.H. Autism: many genes, common pathways? *Cell* **135**, 391-395 (2008).

69. Geschwind,D.H. Autism: many genes, common pathways? *Cell* **135**, 391-395 (2008).

70. Geschwind,D.H. Genetics of autism spectrum disorders. *Trends Cogn Sci.* **15**, 409-416 (2011).

71. Geschwind,D.H., *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69**, 463-466 (2001).

72. Geschwind,D.H., *et al.* The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.* **69**, 463-466 (2001).

73. Goldman,S.E., *et al.* Defining the sleep phenotype in children with autism. *Dev. Neuropsychol.* **34**, 560-573 (2009).

74. Gooneratne,N.S., *et al.* Melatonin pharmacokinetics following two different oral surge-sustained release doses in older adults. *J. Pineal Res.* **52**, 437-445 (2012).

75. Gotham,K., Pickles,A., & Lord,C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J. Autism Dev. Disord.* **39**, 693-705 (2009).

76. Gottesman,I.I. & Gould,T.D. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am. J. Psychiatry* **160**, 636-645 (2003).

77. Grady,B.J., *et al.* Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. *Pac. Symp. Biocomput.*315-326 (2010).

78. Guenole,F., *et al.* Melatonin for disordered sleep in individuals with autism spectrum disorders: systematic review and discussion. *Sleep Med. Rev.* **15**, 379-387 (2011).

79. Hallmayer,J., *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095-1102 (2011).

80. Hammock,E., *et al.* Examining autism spectrum disorders by biomarkers: example from the oxytocin and serotonin systems. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 712-721 (2012).

81. Hanson,D.R. & Gottesman,I.I. The genetics, if any, of infantile autism and childhood schizophrenia. *J. Autism Child. Schizophr.* **6**, 209-234 (1976).

82. Happe,F., Ronald,A., & Plomin,R. Time to give up on a single explanation for autism. *Nat. Neurosci.* **9**, 1218-1220 (2006).

83. Harrington,R.A., *et al.* Serotonin Hypothesis of Autism: Implications for Selective Serotonin Reuptake Inhibitor Use during Pregnancy. *Autism Res.* **6**, 149-168 (2013).

84. Hazen,A. Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of American Society of Civil Engineers* **77**, 1539-1640 (1914).

85. Hemara-Wahanui,A., *et al.* A CACNA1F mutation identified in an X-linked retinal disorder shifts the voltage dependence of Cav1.4 channel activation. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7553-7558 (2005).

86. Holmans,P., *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* **85**, 13-24 (2009).

87. Hotelling,H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417-441 (1933).

88. Hu,V.W., Addington,A., & Hyman,A. Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published GWAS data. *PLoS. One.* **6**, e19067 (2011).

89. Hu,V.W., *et al.* Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Res.* **2**, 78-97 (2009).

90. Hu,V.W., *et al.* Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. *Autism Res.* **2**, 78-97 (2009).

91. Hu,V.W. & Steinberg,M.E. Novel clustering of items from the Autism Diagnostic Interview-Revised to define phenotypes within autism spectrum disorders. *Autism Res.* **2**, 67-77 (2009).

92. Hu-Lince,D., *et al.* The Autism Genome Project: goals and strategies. *Am. J. Pharmacogenomics.* **5**, 233-246 (2005).

93. Hu-Lince,D., *et al.* The Autism Genome Project: goals and strategies. *Am. J. Pharmacogenomics.* **5**, 233-246 (2005).

94. Hubert,L. & Arabie,P. Comparing Partitions. *Journal of Classification* **2**, 193-218 (1985).

95. Hus,V., Gotham,K., & Lord,C. Standardizing ADOS Domain Scores: Separating Severity of Social Affect and Restricted and Repetitive Behaviors. *J. Autism Dev. Disord.*(2012).

96. Hussman,J.P., *et al.* A noise-reduction GWAS analysis implicates altered regulation of neurite outgrowth and guidance in autism. *Mol. Autism* **2**, 1 (2011).

97. IMGSAC A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Hum. Mol. Genet.* **7**, 571-578 (1998).

98. Jamain,S., *et al.* Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nat. Genet.* **34**, 27-29 (2003).

99. Jonsson,L., *et al.* Mutation screening of melatonin-related genes in patients with autism spectrum disorders. *BMC. Med. Genomics.* **3**, 10 (2010).

100. Kalb,L.G., *et al.* Onset patterns prior to 36 months in autism spectrum disorders. *J. Autism Dev. Disord.* **40**, 1389-1402 (2010).

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

102. Kanner,L. Autistic disturbances of affective contact. *Acta Paedopsychiatr.* **35**, 100-136 (1968).

103. Kaufman,L. & Rousseeuw,P. *Finding Groups in Data: An Introduction to Cluster Analysis*(Wiley-Interscience,1990).

104. Kaufmann,W.E., *et al.* Autism spectrum disorder in fragile X syndrome: communication, social interaction, and specific behaviors. *Am. J. Med. Genet. A* **129A**, 225-234 (2004).

105. Kim,S.H. & Lord,C. Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of age. *J. Child Psychol. Psychiatry* **53**, 143-151 (2012).

106. Klei,L., *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 9 (2012).

107. Klei,L., *et al.* Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 9 (2012).

108. Klein,D. Quantification using real-time PCR technology: applications and limitations. *Trends in Molecular Medicine* **8**, 257-260 (2002).

109. Klepstad,P., *et al.* Genetic variability and clinical efficacy of morphine. *Acta Anaesthesiol. Scand.* **49**, 902-908 (2005).

110. Kolevzon,A., Mathewson,K.A., & Hollander,E. Selective serotonin reuptake inhibitors in autism: a review of efficacy and tolerability. *J. Clin. Psychiatry* **67**, 407-414 (2006).

111. Krakowiak,P., *et al.* Sleep problems in children with autism spectrum disorders, developmental delays, and typical development: a population-based study. *J. Sleep Res.* **17**, 197-206 (2008).

112. Krey,J.F. & Dolmetsch,R.E. Molecular mechanisms of autism: a possible role for Ca2+ signaling. *Curr. Opin. Neurobiol.* **17**, 112-119 (2007).

113. Laird,N.M., Horvath,S., & Xu,X. Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* **19 Suppl 1**, S36-S42 (2000).

114. Laird,N.M. & Lange,C. Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* **7**, 385-394 (2006).

115. Laliberte,E. & Legendre,P. A distance-based framework for measuring functional diversity from multiple traits. *Ecology.* **91**, 299-305 (2010).

116. Laliberte,E. & Shipley,B. Measuring functional diversity (FD) from multiple traits, and other tools for functional ecology. R package version 1.0-11. 2011.
Ref Type: Computer Program

117. Laumonnier,F., *et al.* Association of a functional deficit of the BKCa channel, a synaptic regulator of neuronal excitability, with autism and mental retardation. *Am. J. Psychiatry* **163**, 1622-1629 (2006).

118. Le,C.A., *et al.* Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: the ADI-R and the ADOS. *J. Autism Dev. Disord.* **38**, 362-372 (2008).

119. Le,S., Josse,J., & Husson,F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software 25[1], 1-18. 2008.
Ref Type: Computer Program

120. Lemonde,S., *et al.* Association of the C(-1019)G 5-HT1A functional promoter polymorphism with antidepressant response. *Int. J. Neuropsychopharmacol.* **7**, 501-506 (2004).

121. Lesnick,T.G., *et al.* A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS. Genet.* **3**, e98 (2007).

122. Leyfer,O.T., *et al.* Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. *J. Autism Dev. Disord.* **36**, 849-861 (2006).

123. Liu,J., *et al.* A genomewide screen for autism susceptibility loci. *Am. J. Hum. Genet.* **69**, 327-340 (2001).

124. Llerena,A., *et al.* Pharmacogenetics of clinical response to risperidone. *Pharmacogenomics.* **14**, 177-194 (2013).

125. Lopez-Gonzalez,M.A., *et al.* Interaction of melatonin with human lymphocytes: evidence for binding sites coupled to potentiation of cyclic AMP stimulated by vasoactive intestinal peptide and activation of cyclic GMP. *J. Pineal Res.* **12**, 97-104 (1992).

126. Lord,C., Leventhal,B.L., & Cook,E.H., Jr. Quantifying the phenotype in autism spectrum disorders. *Am. J. Med. Genet.* **105**, 36-38 (2001).

127. Lord,C., *et al.* Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J. Autism Dev. Disord.* **19**, 185-212 (1989).

128. Lord,C., Rutter,M., & Le,C.A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24**, 659-685 (1994).

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

130. Lu,A.T., *et al.* QTL replication and targeted association highlight the nerve growth factor gene for nonverbal communication deficits in autism spectrum disorders. *Mol. Psychiatry* **18**, 226-235 (2013).

131. Lu,A.T., *et al.* QTL replication and targeted association highlight the nerve growth factor gene for nonverbal communication deficits in autism spectrum disorders. *Mol. Psychiatry* **18**, 226-235 (2013).

132. Luo,R., *et al.* Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am. J. Hum. Genet.* **91**, 38-55 (2012).

133. Ma,D., *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann. Hum. Genet.* **73**, 263-273 (2009).

134. Ma,D., *et al.* A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann. Hum. Genet.* **73**, 263-273 (2009).

135. Maechler,M., Rouseeuw,P., Struyf,A., Hubert,M., & Hornik,K. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.3. 2012.

Ref Type: Computer Program

136. Makki,N. & Capecchi,M.R. Identification of novel Hoxa1 downstream targets regulating hindbrain, neural crest and inner ear development. *Dev. Biol.* **357**, 295-304 (2011).

137. Malow,B., *et al.* Melatonin for sleep in children with autism: a controlled trial examining dose, tolerability, and outcomes. *J. Autism Dev. Disord.* **42**, 1729-1737 (2012).

138. Mandy,W.P. & Skuse,D.H. Research review: What is the association between the social-communication element of autism and repetitive interests, behaviours and activities? *J. Child Psychol. Psychiatry* **49**, 795-808 (2008).

139. Marazziti,D., *et al.* Presence and characterization of the serotonin transporter in human resting lymphocytes. *Neuropsychopharmacology* **19**, 154-159 (1998).

140. Masana,M.I. & Dubocovich,M.L. Melatonin receptor signaling: finding the path through the dark. *Sci. STKE.* **2001**, e39 (2001).

141. Matson,J.L. & Shoemaker,M. Intellectual disability and its relationship to autism spectrum disorders. *Res. Dev. Disabil.* **30**, 1107-1114 (2009).

142. Matuszek,G. & Talebizadeh,Z. Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. *BMC. Med. Genet.* **10**, 102 (2009).

143. Matuszek,G. & Talebizadeh,Z. Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. *BMC. Med. Genet.* **10**, 102 (2009).

144. Mayr,L.M. & Bojanic,D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **9**, 580-588 (2009).

145. McAllister,A.K. & Van de Water,J. Breaking boundaries in neural-immune interactions. *Neuron* **64**, 9-12 (2009).

146. McPheeters,M.L., *et al.* A systematic review of medical treatments for children with autism spectrum disorders. *Pediatrics.* **127**, e1312-e1321 (2011).

147. Megens,A.A., *et al.* Survey on the pharmacodynamics of the new antipsychotic risperidone. *Psychopharmacology. (Berl.)* **114**, 9-23 (1994).

148. Melke,J., *et al.* Abnormal melatonin synthesis in autism spectrum disorders. *Mol. Psychiatry* **13**, 90-98 (2008).

149. Molloy,C.A., *et al.* Familial autoimmune thyroid disease as a risk factor for regression in children with Autism Spectrum Disorder: a CPEA Study. *J. Autism Dev. Disord.* **36**, 317-324 (2006).

150. Mouridsen,S.E., *et al.* Autoimmune diseases in parents of children with infantile autism: a case-control study. *Dev. Med. Child. Neurol.* **49**, 429-432 (2007).

151. Mulder,E.J., *et al.* Platelet serotonin levels in pervasive developmental disorders and mental retardation: diagnostic group differences, within-group distribution, and behavioral correlates. *J. Am. Acad. Child. Adolesc. Psychiatry* **43**, 491-499 (2004).

152. Nava,C., *et al.* Analysis of the chromosome X exome in patients with autism spectrum disorders identified novel candidate genes, including TMLHE. *Transl. Psychiatry* **2**, e179 (2012).

153. Needleman,L.A. & McAllister,A.K. The major histocompatibility complex and autism spectrum disorder. *Dev. Neurobiol.* **72**, 1288-1301 (2012).

154. Neves-Pereira,M., *et al.* Deregulation of EIF4E: a novel mechanism for autism. *J. Med. Genet.* **46**, 759-765 (2009).

155. O'Dushlaine,C., *et al.* The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics.* **25**, 2762-2763 (2009).

156. O'Roak,B.J., *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585-589 (2011).

157. O'Roak,B.J., *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43**, 585-589 (2011).

158. O'Roak,B.J. & State MW Autism genetics: strategies, challenges, and opportunities. *Autism Res.* **1**, 4-17 (2008).

159. Ozonoff,S., *et al.* Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics.* **128**, e488-e495 (2011).

160. Persico,A.M. & Napolioni,V. Autism genetics. *Behav. Brain Res.* **251**, 95-112 (2013).

161. Peter,B., *et al.* Replication of CNTNAP2 association with nonword repetition and support for FOXP2 association with timed reading and motor activities in a dyslexia family sample. *J. Neurodev. Disord.* **3**, 39-49 (2011).

162. Pinto,D., *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature.* **466**, 368-372 (2010).

163. Piven,J., *et al.* Broader autism phenotype: evidence from a family history study of multiple-incidence autism families. *Am. J. Psychiatry* **154**, 185-190 (1997).

164. Prior,M., *et al.* Are there subgroups within the autistic spectrum? A cluster analysis of a group of children with autistic spectrum disorders. *J. Child Psychol. Psychiatry* **39**, 893-902 (1998).

165. Pritchard,J.K., Stephens,M., & Donnelly,P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

166. Purcell,S., *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559-575 (2007).

167. Rao,N. The clinical pharmacokinetics of escitalopram. *Clin. Pharmacokinet.* **46**, 281-290 (2007).

168. Risch,N., *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493-507 (1999).

169. Rivera-Baltanas,T., *et al.* Serotonin transporter clustering in blood lymphocytes as a putative biomarker of therapeutic efficacy in major depressive disorder. *J. Affect. Disord.* **137**, 46-55 (2012).

170. Roche,A.F., *et al.* Head circumference reference data: birth to 18 years. *Pediatrics* **79**, 706-712 (1987).

171. Rodriguez,I.R., *et al.* Structural analysis of the human hydroxyindole-O-methyltransferase gene. Presence of two distinct promoters. *J. Biol. Chem.* **269**, 31969-31977 (1994).

172. Ronald,A., *et al.* Genetic heterogeneity between the three components of the autism spectrum: a twin study. *J. Am. Acad. Child Adolesc. Psychiatry* **45**, 691-699 (2006).

173. Ronald,A., Happe,F., & Plomin,R. The genetic relationship between individual differences in social and nonsocial behaviours characteristic of autism. *Dev. Sci.* **8**, 444-458 (2005).

174. Roses,A.D., *et al.* Complex disease-associated pharmacogenetics: drug efficacy, drug safety, and confirmation of a pathogenetic hypothesis (Alzheimer's disease). *Pharmacogenomics. J.* **7**, 10-28 (2007).

175. Rossignol,D.A. & Frye,R.E. Melatonin in autism spectrum disorders: a systematic review and meta-analysis. *Dev. Med. Child. Neurol.* **53**, 783-792 (2011).

176. Sanders,S.J., *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).

177. Sanders,S.J., *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature.* **485**, 237-241 (2012).

178. SCHAIN,R.J. & FREEDMAN,D.X. Studies on 5-hydroxyindole metabolism in autistic and other mentally retarded children. *J. Pediatr.* **58**, 315-320 (1961).

179. Schellenberg,G.D., *et al.* Evidence for multiple loci from a genome scan of autism kindreds. *Mol. Psychiatry* **11**, 1049-60, 979 (2006).

180. Scott,L.J. & Dhillon,S. Risperidone: a review of its use in the treatment of irritability associated with autistic disorder in children and adolescents. *Paediatr. Drugs.* **9**, 343-354 (2007).

181. Sevin,J.A., *et al.* Empirically derived subtypes of pervasive developmental disorders: a cluster analytic study. *J. Autism Dev. Disord.* **25**, 561-578 (1995).

182. Shao,Y., *et al.* Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *Am. J. Hum. Genet.* **72**, 539-548 (2003).

183. Shao,Y., *et al.* Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *Am. J. Hum. Genet.* **72**, 539-548 (2003).

184. Shao,Y., *et al.* Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *Am. J. Hum. Genet.* **70**, 1058-1061 (2002).

185. Sheldon,S.H. Pro-convulsant effects of oral melatonin in neurologically disabled children. *Lancet.* **351**, 1254 (1998).

186. Sie,L., Loong,S., & Tan,E.K. Utility of lymphoblastoid cell lines. *J. Neurosci. Res.* **87**, 1953-1959 (2009).

187. Siegel,B., *et al.* Empirically derived subclassification of the autistic syndrome. *J. Autism Dev. Disord.* **16**, 275-293 (1986).

188. Silverman,J.M., *et al.* Symptom domains in autism and related conditions: evidence for familiality. *Am. J. Med. Genet.* **114**, 64-73 (2002).

189. Simonoff,E., *et al.* Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *J. Am. Acad. Child. Adolesc. Psychiatry* **47**, 921-929 (2008).

190. Sivertsen,B., *et al.* Sleep problems in children with autism spectrum problems: a longitudinal population-based study. *Autism* **16**, 139-150 (2012).

191. Slominski,R.M., *et al.* Melatonin membrane receptors in peripheral tissues: distribution and functions. *Mol. Cell Endocrinol.* **351**, 152-166 (2012).

192. Snow,A.V., Lecavalier,L., & Houts,C. The structure of the Autism Diagnostic Interview-Revised: diagnostic and phenotypic implications. *J. Child Psychol. Psychiatry* **50**, 734-742 (2009).

193. Sparrow,S., Cicchetti,D., & Balla,D. Vineland Adaptive Behavior Scales, Second Edition. 2005. Minneapolis, MN, Pearson Assessments.
Ref Type: Pamphlet

194. Sparrow,S., Balla,D., & Cicchetti,D. Vineland Adaptive Behavior Scales: Survey form manual. 1984. Circle Pines, MN, American Guidance Service.
Ref Type: Pamphlet

195. Sparrow,S. & Cicchetti,D. Diagnostic uses of the Vineland Adaptive Behavior Scales. *J. Pediatr. Psychol.* **10**, 215-225 (1985).

196. Sparrow,S., Cicchetti,D., & Balla,D. VINELAND ADAPTIVE BEHAVIOR SCALES, SECOND EDITION (VINELAND-II), 2005. 2005. NCS Pearson, Inc.
Ref Type: Pamphlet

197. Spiker,D., *et al.* Behavioral phenotypic variation in autism multiplex families: evidence for a continuous severity gradient. *Am. J. Med. Genet.* **114**, 129-136 (2002).

198. Splawski,I., *et al.* Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* **119**, 19-31 (2004).

199. Splawski,I., *et al.* CACNA1H mutations in autism spectrum disorders. *J. Biol. Chem.* **281**, 22085-22091 (2006).

200. Statacorp. Stata Statistical Software. [11]. 2009. College Station, TX, Statacorp LP.
Ref Type: Computer Program

201. Stevens,M.C., *et al.* Subgroups of children with autism by cluster analysis: a longitudinal examination. *J. Am. Acad. Child Adolesc. Psychiatry* **39**, 346-352 (2000).

202. Suda,S., *et al.* Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol. Autism* **2**, 14 (2011).

203. Sudhof,T.C. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature.* **455**, 903-911 (2008).

204. Talebizadeh,Z., Arking,D., & Hu,V. A Novel Stratification Method in Linkage Studies to Address Inter- and Intra-Family Heterogeneity in Autism. *PLoS. Genet.* **8**, e67569 (13 A.D.).

205. Talebizadeh,Z., Arking,D., & Hu,V. A Novel Stratification Method in Linkage Studies to Address Inter- and Intra-Family Heterogeneity in Autism. *PLoS. Genet.* **8**, e67569 (2013).

206. ten Asbroek,A.L., *et al.* Polymorphisms in the large subunit of human RNA polymerase II as target for allele-specific inhibition. *Nucleic. Acids Res.* **28**, 1133-1138 (2000).

207. Tetko,I.V., Livingstone,D.J., & Luik,A.I. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. In and Comput. Sci.* **35**, 826-833 (1995).

208. Tischfield,M.A., *et al.* Homozygous HOXA1 mutations disrupt human brainstem, inner ear, cardiovascular and cognitive development. *Nat. Genet.* **37**, 1035-1037 (2005).

209. Tuchman,R. & Cuccaro,M. Epilepsy and autism: neurodevelopmental perspective. *Curr. Neurol. Neurosci. Rep.* **11**, 428-434 (2011).

210. Turner,S. Annotated Manhattan plots and QQ plots for GWAS using R, Revisited. 2011. Nature Precedings.
Ref Type: Online Source

211. Vassilev,L.T., *et al.* In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science.* **303**, 844-848 (2004).

212. Veatch,O.J., *et al.* Identification of Genetically Meaningful Phenotypic Subgroups in Autism Spectrum Disorders. *Under Review*(2013).

213. Veenstra-VanderWeele,J., Christian,S.L., & Cook,E.H., Jr. Autism as a paradigmatic complex genetic disorder. *Annu. Rev. Genomics Hum. Genet.* **5**, 379-405 (2004).

214. Veenstra-VanderWeele,J. & Cook,E.H., Jr. Molecular genetics of autism spectrum disorder. *Mol. Psychiatry* **9**, 819-832 (2004).

215. Vernes,S.C., *et al.* A functional genetic link between distinct developmental language disorders. *N. Engl. J. Med.* **359**, 2337-2345 (2008).

216. Virkud,Y.V., *et al.* Familial aggregation of quantitative autistic traits in multiplex versus simplex autism. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **150B**, 328-334 (2009).

217. Voineagu,I., *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature.* **474**, 380-384 (2011).

218. Wang,K., Li,M., & Bucan,M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278-1283 (2007).

219. Wang,L., *et al.* Sequencing ASMT identifies rare mutations in Chinese Han patients with autism. *PLoS. One.* **8**, e53727 (2013).

220. Wang,L.W., Tancredi,D.J., & Thomas,D.W. The prevalence of gastrointestinal problems in children across the United States with autism spectrum disorders from families with multiple affected members. *J. Dev. Behav. Pediatr.* **32**, 351-360 (2011).

221. Weiss,L.A., *et al.* A genome-wide linkage and association scan reveals novel loci for autism. *Nature.* **461**, 802-808 (2009).

222. Weiss,L.A., *et al.* Sodium channels SCN1A, SCN2A and SCN3A in familial autism. *Mol. Psychiatry* **8**, 186-194 (2003).

223. Whitehouse,A.J., *et al.* CNTNAP2 variants affect early language development in the general population. *Genes Brain Behav.* **10**, 451-456 (2011).

224. Wickham,H. ggplot2: an Implementation of the Grammar of Graphics. http://cran.r-project.org/web/packages/ggplot2.[R package Version 0.7]. 2008.
Ref Type: Computer Program

225. Wiggins,L.D., *et al.* Support for a dimensional view of autism spectrum disorders in toddlers. *J. Autism Dev. Disord.* **42**, 191-200 (2012).

226. Wilens,T.E., *et al.* Fluoxetine pharmacokinetics in pediatric patients. *J. Clin. Psychopharmacol.* **22**, 568-575 (2002).

227. Williams,C.A., Dagli,A., & Battaglia,A. Genetic disorders associated with macrocephaly. *Am. J. Med. Genet. A* **146A**, 2023-2037 (2008).

228. Williams,P.G. & Hersh,J.H. Brief report: the association of neurofibromatosis type 1 and autism. *J. Autism Dev. Disord.* **28**, 567-571 (1998).

229. Yaspan,B.L., *et al.* Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.* **129**, 563-571 (2011).

230. Yonan,A.L., *et al.* A genomewide screen of 345 families for autism-susceptibility loci. *Am. J. Hum. Genet.* **73**, 886-897 (2003).