Distribution Agreement

Signature: ________________________ Date: 25/Apr/2024

LARGE EXOME SEQUENCING STUDY ANALYZING X-LINKED VARIATION IN AUTISM SPECTRUM DISORDER

By

Nicholas Weaver
Degree to be awarded: Master of Public Health

Department of Epidemiology

____
Dr. David Cutler, PhD
Committee Chair

____
Dr. Yan Sun, PhD, M.S.
Committee Member

LARGE EXOME SEQUENCING STUDY ANALYZING X-LINKED VARIATION IN AUTISM SPECTRUM DISORDER

By

Nicholas Weaver

Bachelor of Science, Molecular and Cellular Biology
University of Illinois Urbana-Champaign
2022

Thesis Committee Chair: Dr. David Cutler, PhD
Thesis Committee Member: Dr. Yan Sun, PhD, M.S.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University in
partial fulfillment of the requirements for the degree of
Master of Public Health in Epidemiology
2024

Abstract

LARGE EXOME SEQUENCING STUDY ANALYZING X-LINKED VARIATION IN AUTISM SPECTRUM DISORDER

By Nicholas Weaver

Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by a spectrum of challenges, including difficulties with social communication, repetitive behaviors, limited interests, and sensory sensitivities. Research shows a higher prevalence of ASD in males compared to females, with estimates ranging from 2:1 to 5:1. One area of particular interest in ASD research is the role of the X chromosome. We hope to identify variants strongly associated with the ASD and explore whether these variants have a differential impact on ASD risk in males and females.

Methods

A total of approximately 74,000 individuals were analyzed, including around 20,000 individuals with ASD. Likelihood ratio tests corresponding to 3 different models of genetic effect: additive, recessive, recessive lethal were analyzed and p-values were FDR corrected. QQ plots were ran to find genetic variants meeting QC criteria and FDR thresholds. Penetrance of variants were calculated via liability scores produced by Quantitative X-Linked Transmitted and De Novo Analysis (QXL-TADA), and odds ratios (ORs) calculated via penetrances.

Results

In this analysis 11 genes were associated with one of our three models, and therefore associated with ASD. Among these 11 genes include variants of additive effect (ARHGEF9, $p < 1e-8$), recessive effect (DGAT2L6, $p < 0.001$) and lethal effect (DDX3X, $p < 1e-8$). The overall penetrances of the variants associated with ASD are often higher in males than in females. However, odds ratios (ORs) tell a different story with the rare homozygous females (A2A2) often having a higher OR, and thus higher risk of ASD than the rare hemizygous males (A2(-)).

Conclusion

Additive effect variants accounted for six genes associated with ASD, recessive effect variants accounted for four genes, and lethal effect variants accounted for one gene associated with ASD. The majority of genes were associated with neural pathways, and remaining genes were involved in transcriptional or metabolic regulation. Interestingly, males are not always at a higher risk for ASD when carrying rare variants. This study sheds light on the prevalence differences of ASD attributed to the X chromosome, variants associated with ASD, and attempts to reconcile some of the diagnostic disparities we see between males and females.

LARGE EXOME SEQUENCING STUDY ANALYZING X-LINKED VARIATION IN AUTISM SPECTRUM DISORDER

By

Nicholas Weaver
Bachelor of Science, Molecular and Cellular Biology
University of Illinois Urbana-Champaign
2022

Thesis Field Advisor: Dr. David Cutler
Thesis Faculty Advisor: Dr. Yan Sun

Thesis submitted to the Faculty of the Rollins School of Public Health of Emory University in partial fulfillment of the requirements for the degree of Master of Public Health (MPH) in Epidemiology

Department of Epidemiology
Emory University, Rollins School of Public Health

Department of Human Genetics
Emory University, School of Medicine

2024

SUMMARY

This study analyzes X-linked variation in Autism Spectrum Disorder (ASD) with data from nearly 74,000 case-control and family probands and identifies 11 genes that are significantly associated with ASD. We characterize these genes by their inferred mode of inheritance: additive, recessive, or recessive lethal via likelihood ratio tests, and subsequently calculate the penetrances and odds ratios of genes associated with ASD. Of the 11 associated genes, six genes appear to have neurological specific effects, three genes likely have broader effects on transcriptional regulation and two genes were associated with other metabolic cellular processes. This study gives in-depth analysis of the X chromosome and its rare genetic variants associated with ASD and attempts to use this information to reconcile some of the diagnostic disparities that exist between male and female diagnostic rates for ASD.

INTRODUCTION

The etiology, biology, and genetic architecture of Autism Spectrum Disorder (ASD) has been debated significantly in the fields of molecular biology, environmental health, and genetic epidemiology throughout the last few decades. ASD is characterized by a wide range of impairments such as: deficits in social communication, repetitive behaviors, limited interests, and a sensitivity to external stimuli like certain fabrics or light. ASD, a polygenetic disorder, effects numerous different genes throughout the human genome, and debates circulate whether environmental stimuli and genetic predisposition or solely genetic makeup is responsible for the etiology of ASD **(1-3)**. As of 2024, there is no known cure, and most medical treatments involve treating the symptoms of ASD rather than investigating the genetic underpinnings of the disorder as a whole. Numerous studies have attempted to categorize the genetic component of ASD through identifying novel risk genes **(1-2)** associated with ASD in numerous genome-wide-association-studies (GWAS), whole-exome-sequencing (WES) studies, and autosome data in cohort and case-control studies. However, the genetic composition of the X chromosome is often wholly left out of consideration when searching for genes that implicate ASD in these samples **(2)**. An article published by Maenner et al. **(4)** found that the male: female prevalence of ASD is around 4.3-1, while other studies site that difference is somewhere between 2:1 and 5:1 **(5,6)**. This type of difference, corroborated by numerous studies, could be interpreted as non-random association between males and a higher likelihood for ASD diagnosis. Numerous theories have been proposed to explain the difference in the male: female ratio of ASD, those being the Extreme Male Brain Theory (EMB), Female Protective Theory, and the Female Autism Phenotype theory. EMB states that the neurotomical make-up of those with ASD more closely resembles the male brain than the female brain, thus making males more predispositioned to ASD, corroborated by the fact that those with ASD often have more interests and behavioral patterns more closely aligned with maleness such as a more significant interest in things and objects **(7)**. Likewise, the Female Protective Theory **(8-9)** states that since females have two X chromosomes, the genes contained for brain development are protected if there were to be a deficit in one of the X chromosomes, although this theory would have to explain the effect of dosage compensation for the majority of genes on the X chromosome **(10)**. Lastly, the Female Autism Phenotype theory **(9)** calls into question the diagnostic criteria and how ASD is manifested differently in females than in males. This theory states that females with ASD manifest their symptoms differently with regards to social skills and overall friendship quality than males with ASD. Since the structure of ASD is laid out to diagnose males, it is thought that the females with ASD get underdiagnosed due to them scoring similarly to non-autistic boys on social and friendship criteria of an ASD diagnostic test **(11)**. Other accounts interpret this phenomenon more simply, essentially saying females, as a whole, are just more difficult to diagnose and regard this theory as more of an observation rather than a gap in the literature **(12)**. However, diagnostic criteria can only explain away some of the clinical manifestations of 4.3-1, male: female difference for ASD shown in numerous studies.

Genetics of the X chromosome

Genetically, the sex chromosomes of males are distinct from their female counterparts due to the hemizygosity of the X chromosome. Therefore, the penetrance of many X linked disorders is near 100% for males with an effected genetic variant but differs greatly in females due to the variable expressivity and mosaic expression of the X chromosome throughout all the cells in the body **(13)**. There are approximately 800 protein coding genes on the X chromosome, many of which code for important biological process such as development/growth, or neurological function **(14,15**) shown by deficits presented in disorders such as Fragile X syndrome associated with severe intellectual disability and a condition that is often associated with ASD diagnosis. Since ASD is a polygenic disorder effecting numerous genes throughout the entire genome, and effect males at 4-5 times the rate of females; it would make sense for studies to include the X chromosome in hopes of explaining the reasoning behind an increased susceptibility to ASD.

Initial WES Study

Studies to date have been qualitative in nature **(8)**, genome-wide but excluded the X chromosome **(1-2)** or defined epigenetic regions of the genome including differentially methylated regions (DMRs) **(3)**. A study by Satterstrom et al **(2)**., published a study analyzing all 22 pairs of autosomes using large scale WES comprising of 35,584 total samples, identified 102 risk genes to be associated with ASD. Of the 102 genes analyzed, 49 of the are associated with severe developmental delay via de novo mutations. De novo mutations can be classified as any mutations not directly inherited from either parent. De novo mutations are sporadic, random, and can cause severe deficits in the health of an individual; especially being common in many forms of ASD. The remaining 53 risk genes have been shown to be common in those diagnosed with ASD but are not always shown in those with developmental disorders. Many of genes discovered by Satterstrom et al. were shown to be expressed in cell types consistent with neurological function, specifically in the excitatory and inhibitory neural signals, and the communication within those cell types. The method used to analyze this amount of data for assessing risk genes was by dividing genetic variants into 7 classes based on their missense badness, PolyPhen-2, constraint (MPC) score, and their probability loss-of-function intolerance (pLI) score. This system allowed for the researchers to assess multiple levels of severity based on the deviation from the wild-type (WT) version of the risk genes being analyzed based on the level of functional deficit. The tier with the most functional deficit were mutations that induced protein-truncating variants (PTVs), in which the protein is truncated, or shortened, and shows a limited to null effect inside the cell. These deficits encompassed 3 tiers (≥0.995, 0.5–0.995, 0–0.5) with ≥0.995 being the most severe. The next 3 tiers represent missense mutations, which are mutations that often make a full transcript, but with one or more amino acids changed, therefore limiting function. Those tiers are represented by the missense badness, PolyPhen-2, constraint (MPC) score **(16)** which considers position and how likely the variant is to cause decreased protein function. This scale is typically scored on a scale of 0-≤ 2 with 3 tiers as well (≤2, 1-2, 0). The final tier is for synonymous (SYN) variants which are point mutations that do not change the genetic code and therefore, do not change the protein structure or function. Through this system, researchers are able to analyze, prioritize, and identify the genetic variants that most greatly impede biological processes. Thereby, prioritizing variants for further research in the development of therapies and pharmaceuticals that could one day lead to more beneficial health outcomes.

The previous study points out the X chromosome was not included in the WES analysis, and further research needs to be done on the X chromosome as it pertains to ASD and neuronal developmental disorders (NDD). Some studies have suggested that there are far more genes pertaining to complex regulatory and brain function on the X chromosome **(17)**, with nearly 20% of these genes being expressed in neuronal communication and cognitive functions **(18**). It is worth noting that the Satterstrom et al., article showed the effect sizes for various genetic variants between males and females, and these effect sizes were not significantly or statistically different from each other. Pertaining to the disparity between sexes for ASD diagnosis, it is clear whatever is leading to the increased diagnosis of males is not being caused by any rare variants on the autosomes. In this study, we will be analyzing the PTVs of the X chromosome. We will exclude some of the tiers used in the Satterstrom paper because we are mostly concerned with variants that confer loss-of-function (PTV) and the associated liabilities for ASD.

Unlocking the potential risk genes located within the X chromosome might be crucial to understanding the sex difference in ASD and unlock new therapeutic interventions or a field of research dedicated to curtailing the severe deficits associated with ASD diagnosis.

Analyzing the X Chromosome

With significant sex disparities between males and females related to ASD, the importance of analyzing the X chromosome may reveal genes that contribute to the increased diagnosis rate of

ASD in males. A few genes on the X chromosome have already been well studied and related to genetic deficits in both males and females. Fragile X mental retardation 1 (FMR1) is a gene contained within the X chromosome that is involved in neural synapse formation **(19)** and is defective within those with Fragile X syndrome (FXS). This disorder is often graded, showing deficiencies along a spectrum of mild impairment and full mental retardation, and often occurs around 1.5x higher in males **(20)**. Similarly, the mutation of the MECP2 gene is known to confer Rett Syndrome in effected individuals. MECP2, much like FMR1, has also been shown to be instrumental in proper CNS functioning among healthy individuals, and plays an immediate role in spontaneous neurotransmission and short-term synaptic plasticity **(21)**. Moreover, unlike Fragile X Syndrome which occurs more readily in males, Rett Syndrome is seen almost exclusively in females **(22)**. This could be related to a host of different reasons; however, this phenomenon is probably explained by some factors related to hemizygous lethality of the specific MECP2 mutation causing males to die in-utero. Nevertheless, two of the most well studied disorders related to the X chromosome are wholly associated with neurological deficits; leading many to speculate what role the sex chromosomes play in genetic disorders related to neurological function.

X-linked Variation in ASD

In terms of this study, we will analyze 74,000 case-control and family probands comprising of 16,000 males and 4,000 females with ASD, 8,000 males and 6,000 females without ASD, along with and additional 40,000 parents without ASD. We are interested in analyzing the effects of variants related to ASD specifically on the X chromosome. Ideally, we are hoping to gather information on which specific genes are significantly associated with ASD, and if there are any differing rates of ASD for males and females with regards to variants of those genes. Since the X chromosome is vastly important in terms of brain function, and genes located within the X chromosome are extensively expressed in the human brain **(23)**, subsequent analysis should be able to locate a host of genetic variants that confer some susceptibility to ASD, or at the very least, narrow the field of ASD research to specific areas of the genome known to be effected in those who meet the threshold for ASD diagnosis. It is worth noting that while this study will be exclusively on the sex chromosomes, the Y chromosome will not be analyzed simply due to the lack of genes and the fact that the Y chromosome is exclusively male-specific unlike the X chromosome.

METHODS

Samples

Samples were analyzed across four datasets: 1) the ASC v17 dataset, containing ASC sequencing batches 1-14, as well the Simons Simplex Collection (SSC), 2) a separate dataset containing ASC sequencing batches 15 and 16, 3) the Simons Foundation Powering Autism Research for Knowledge (SPARK) Pilot dataset, and 4) the SPARK 2019 release of approximately 27,000 samples ("SPARK 27k"). All four contained family-based trio data, and the ASC v17 VCF also contained Swedish PAGES case-control samples. In addition, we incorporated counts from autism cases and controls from the Danish iPSYCH cohort **(2)**. Overall, these are the same datasets used in Fu et al. 2022 **(24)**, although this paper uses PAGES samples from the v17 VCF (rather than lifting over data from the same samples published in Satterstrom et al 2020, **(2)** and does not include the 458 probands and 101 siblings incorporated into Fu et al 2022 from published data (due to lack of inherited variant information)

Samples by Dataset

	Family				Case-Control			
	Male		Female		Male		Female	
	Probands	Siblings	Probands	Siblings	Case	Control	Case	Control
ASC v17	6025	1158	1265	1190	517	1845	210	1741
ASC B15-B16	223	6	56	5				
SPARK Pilot	376		89					
SPARK 27k	5219	1554	1324	1478				
iPSYCH					3730	3373	1133	1629
Totals	11843	2718	2734	2673	4247	5218	1343	3370

Sequencing

Production of the datasets analyzed in this study using the genome analysis toolkit (GATK, for which we cite "Genomics in the Cloud: Using Docker, GATK, and WDL in Terra") has been described in Fu et al. (2022**) (24)**. Briefly, ASC and SSC samples were processed by aligning sequence read data to the hg38 reference genome. Variants were first called individually using local realignment by HaplotypeCaller in gVCF mode and were then called jointly using GenotypeGVCFs. Variant quality score recalibration (VQSR) was run on the joint dataset to estimate variant call accuracy. SPARK Pilot bam files aligned to the hg19 genome were downloaded from SFARI, realigned to the hg38 reference genome, and processed using the same pipeline as AC+SSC data. For the larger SPARK 27k release, individual gvcf files produced by GATK were downloaded from SFARI, variants were called jointly, and VQSR was run on the resulting dataset.

Dataset processing and QC

The same working datasets were used as described in Fu et al. 2022 **(24)**. Briefly, Hail 0.2 (https://hail.is) was used to process the four datasets individually. Low-complexity regions (using https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs38.bed.gz) were dropped, and variants were assigned specific genes and consequence values by annotation with the Variant Effect Predictor (VEP) **(25).** Genotypes were required to have a minimum depth of 10, except for male hemizygous regions, where a minimum of 7 was required. Genotypes were also filtered if the depth exceeded 1000. Homozygous reference calls were required to have a genotype quality (GQ) at least 25, and heterozygous and homozygous variant calls were required to have a phred-scaled likelihood of the call being homozygous reference (PL[HomRef]) of at least 25. Additionally, heterozygous calls in male hemizygous regions and any calls on the Y chromosome in females were filtered.

Genotypes were further filtered if (1) the allele balance (# reads supporting the alternate allele/depth) of a heterozygous call was below 0.25 (2) the probability of the allele balance <1e-8, assuming a binomial distribution with mean 0.5 or (3) the number of informative reads supporting a heterozygous call (counting reads supporting either the reference or alternate allele) or homozygous call (counting reads supporting the alternate allele) was less than 90% of the depth. Finally, variants with a call rate < 10% or a Hardy-Weinberg p-value < 1e-12 were also dropped.

De novo variant calling and quality control

De novo variants were called using Hail's "de-novo" function for genotypes with GQ>=25, using variant frequencies from the non-neuro subset of gnomAD GRCh38 exomes v2.1.1 (gs://gnomad-public/release/2.1.1/liftover_grch38/ht/exomes/gnomad.exomes.r2.1.1.sites.liftover_grch38.ht) as previously described **(24)**.

Case-control variants and QXL-TADA

Briefly, we have extended the TADA modeling framework to account for the inherent differences of the X in a version we are calling QXL-TADA, for quantitative, X-linked, TADA. It is "quantitative" because we have moved to the "quantitative / liability" scale for effect estimation (away from relative risk scales) to better account for male / female differences. It is "X-linked" because it includes multiple disease models to account for complications the X introduces. At the moment, QXL-TADA has no disease prior (it is effectively a pure-penalized likelihood approach), because the years of evidence gathered from the autosomes to build the original TADA prior is not obviously extensible to the X. QXL-TADA is a gene based test, and different variants of the same "class" are assumed to have the same effect, and modeled as a single allele with frequency the sum of the individual allelic variants.

QXL-TADA models the distribution of disease risk and prevalence in a population using the mixed model of inheritance of Morton and MacLean **(26).** In this framework, each individual has some underlying quantitative risk, or liability of developing ASD. If the many genetic and environmental factors contributing to an individual's liability are independent and additive on some scale, then, by the central limit theorem, liability is approximately normally distributed in the population. Individuals with liabilities above some threshold are affected with the disease, and individuals with liabilities below that threshold are not diagnosed as having ASD.

In a genetic association study, we are interested in whether or not variation at a particular locus affects liability for a particular disease. For a biallelic gene, the population liability distribution can be dissected into three curves, representing liability distributions for individuals with each of the three gene genotypes: A1A1, A1A2, A2A2, with A1 representing the major allele, and A2 representing the effect/minor allele. Under the null hypothesis, variation at the gene is not associated with disease risk, and thus, the fraction of individuals with liability greater than the threshold (the genotype-specific penetrance) is the same for all three genotypes. In a more interesting scenario, one allele is associated with increased risk of disease, and the penetrances of the three genotypes differ.

We use the genotype distributions in parent-child trios, case-control data, or any combination therefore, to detect differences in mean liability by gene genotype. Observed case-parent, for instance, genotype counts in a data set are modeled as independent draws from a multinomial distribution defined by several parameters to account for population structure / demographic confounding. In addition to the Null model, we explore three alternative models. We numerically calculate maximum likelihood estimates for the free parameters in the Null and disease models, compare them and assign p-values via likelihood ratio tests. In addition to the additive model (Model 1), we examine a completely recessive model (Model 2), and model of homozygous lethality (Model 3).

R Programming

After initial QXL-TADA analysis, text files were analyzed via R programming for the synonymous (SYN) site variants and the protein-truncating variants (PTV) associated with ASD. For all 726 genes analyzed on the X chromosome, null p-values were given based on the association of disease model of best fit and subsequent association with ASD. As in most genetic association studies, null p-values underwent a False-Discovery rate (FDR) correction for multiple testing via Benjamini-Hochberg (BH) procedure. P-values of both SYN variants and PTV variants were analyzed via QQ plots (qqman package) with FDR-corrected thresholds of <0.01, <0.05, and <0.10 displayed. The BH correction effectively nullified all synonymous variants and allowed us to focus on the 15 genes in the PTV dataset that meant threshold criteria, and thus, should be further analyzed for association with ASD.

As stated above, 15 genes were defined for further analysis as we are interested in the penetrances conferred by each of their rare variants that most readily confer ASD. As defined in the QXL-TADA procedure, liability scores were generated with {0} conferring no disease risk, {0>} conferring a protective effect against ASD, and {0<} conferring some risk for increased prevalence of disease. We are interested in calculating to what degree these liability scores meet threshold for ASD, and how variation between alleles of homozygotes, heterozygous, and hemizygous males associate with manifestation of ASD. The modes in which we analyzed this were in two steps. First, normal distributions curves were produced with ggplot2 tools on R programming using the liability scores as means for each of the associated curves for males and females. A total of 3 curves were produced for females (A1A1, A1A2, A2A2), and a total of 2 curves were produced for males (A1(-), A2(-)). Since the QXL-TADA analysis produces liability scores by allelic variation, each liability score is plotted as a mean on each of the curves that follow a normal distribution. As stated, those diagnosed with ASD should be above a certain threshold. In this study, the subsequent threshold used in this study for ASD diagnosis used was 2.510 for females and 1.885 for males, corresponding to the population prevalence in males (0.0297) and the population prevalence in females (0.0060); the following thresholds were plotted on the graphs. Areas under the curve, past threshold, were then shaded to visualize the distribution of the allelic variants that confers ASD in the 15 genes that passed the FDR-correction criteria. Effectively, the shading under the curves indicates the penetrance of ASD or the likelihood that carrying that specific genotype results in ASD diagnosis. Secondly, calculations were done via inverse pnorm function on R. This procedure calculates the penetrance given the mean (liability score) and the threshold for ASD diagnosis.

$$\text{standardized threshold} = \frac{(Threshold - mean)}{SD}$$

$$\text{Penetrance} = 1 - pnorm(standardized\_threshold)$$

Excel formulas: Calculation of Odds Ratios (ORs)

Furthermore, after penetrance was calculated for each of the 15 genes; odds ratios (ORs) were calculated on excel using the A1A1 penetrance as the reference group for females, and A1(-) for the males. These calculations result in 2 ORs for females (A1A2, A2A2) and 1 OR for males (A2(-)). Calculation below:

Female OR- A1A2: $= \frac{\left(\frac{P(A1A2)}{1-P(A1A2)}\right)}{\left(\frac{P(A1A1)}{1-P(A1A1)}\right)}$	Female OR- A2A2: $= \frac{\left(\frac{P(A2A2)}{1-P(A2A2)}\right)}{\left(\frac{P(A1A1)}{1-P(A1A1)}\right)}$	Male OR- A2(-): $= \frac{\left(\frac{P(A2)}{1-P(A2)}\right)}{\left(\frac{P(A1)}{1-P(A1)}\right)}$

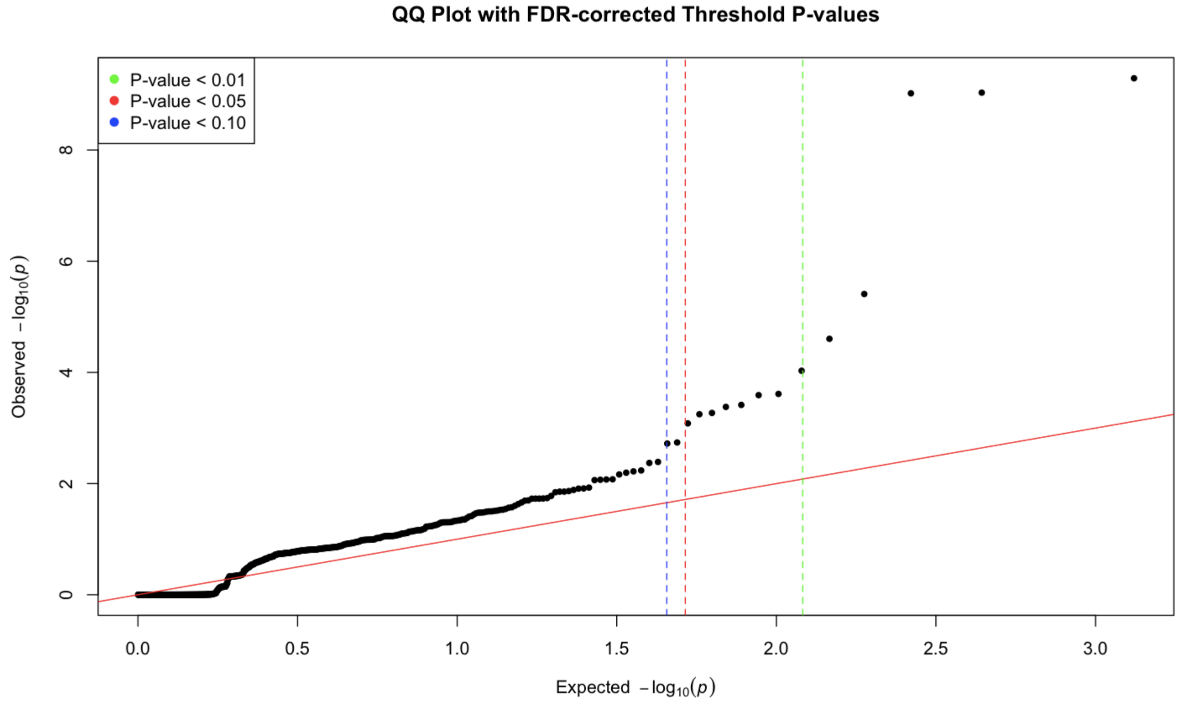
P*= Penetrance

Odds ratios are calculated to quantify the risk of the allelic variants for ASD. All comparisons were to homozygous or hemizygous reference alleles whose variants did not confer any increased risk of ASD above chance. In other words, all variants analyzed whose genes contained only reference alleles did not show any increase in ASD risk, that wouldn’t otherwise be seen due to chance within the population. However, due to the additive, recessive, and lethal nature of the genetic variants for the rare alleles, we should observe different odds ratios for heterozygotes and rare allele homozygotes/hemizygotes in the overall analysis.

All definitions of genes and subsequent functions are provided through the NIH National Library of medicine NBCI gene search engine: https://www.ncbi.nlm.nih.gov/gene/

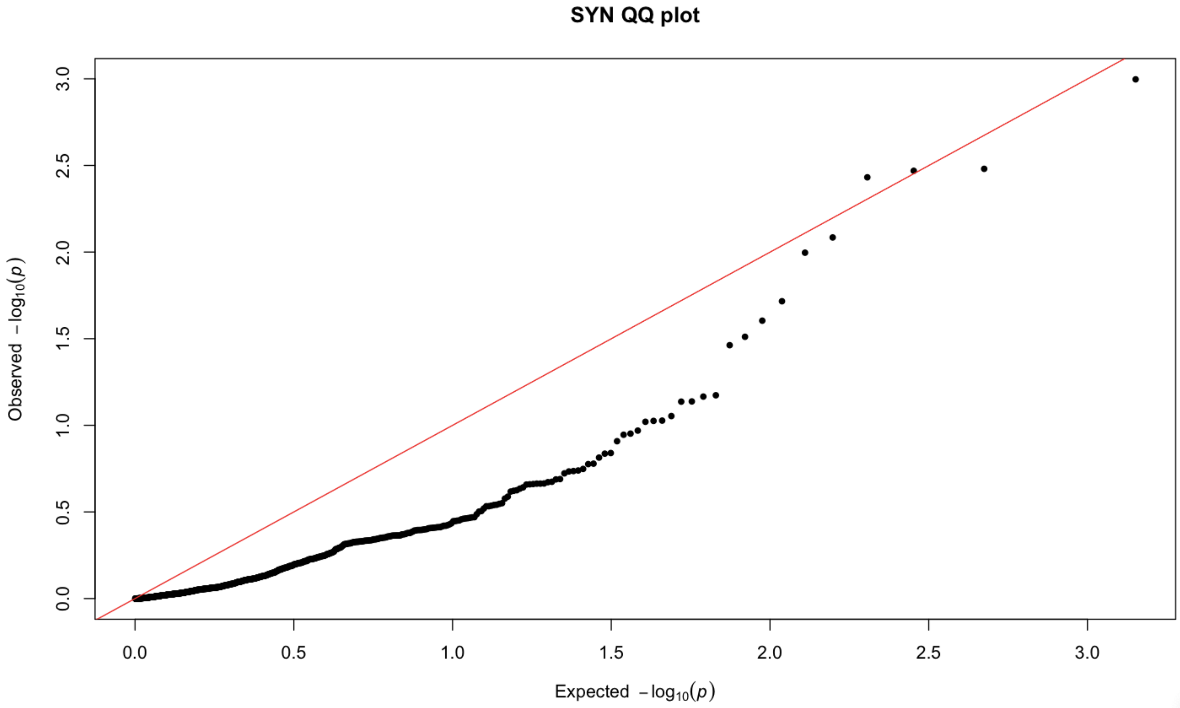
RESULTS

PTV QQ Plot



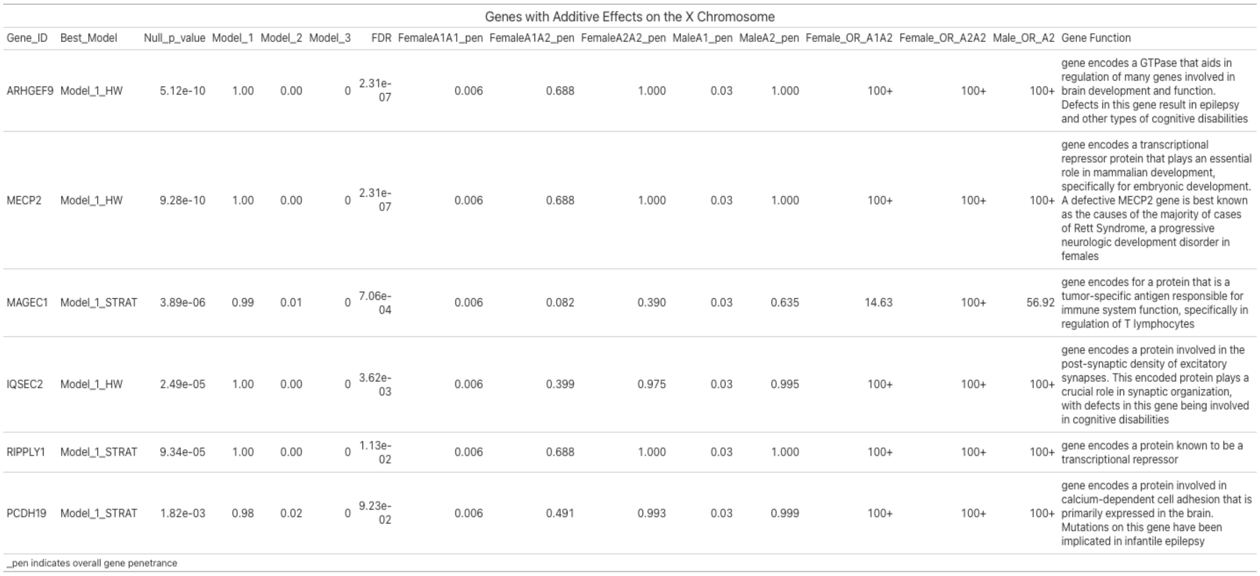
Observed/Expected of the null p-values for the likelihood ratio test were plotted via QQ plot for the protein-truncating variants (PTVs) thought to be associated with ASD. FDR thresholds were plotted vertically at <0.10, <0.05, and <0.01, respectively. Null p-values post-BH correction yielded 15 genes that met FDR criteria. All 15 genes are shown to meet FDR thresholds as well as be above the diagonal line indicating that on the -$\log_{10}$(p-value) scale the observed genes have a larger than expected p-value compared to the normal distribution under the null. In this case, that indicates very small p-values associated with the null model. This allows us to reject the null model distribution and analyze these genes further for model of best fit: additive, recessive, lethal, as it pertains to ASD diagnosis.

SYN QQ Plot



Observed/Expected of the null p-values for the likelihood ratio test were plotted via QQ plot for the synonymous variants thought to be associated with ASD. It is shown that the observed values, as it pertains to the null model fitting the genes of interest, are below what is expected in a normal distribution. Unlike our PTV plot, no SYN variants were discovered below the null p-value of <0.10 post-FDR correction, and overall p-values appear considerably deflated relative to the null model. There are no synonymous variants to have a likely effect on the incidence of ASD.

Table 1- Additive Effect Genes associated with ASD



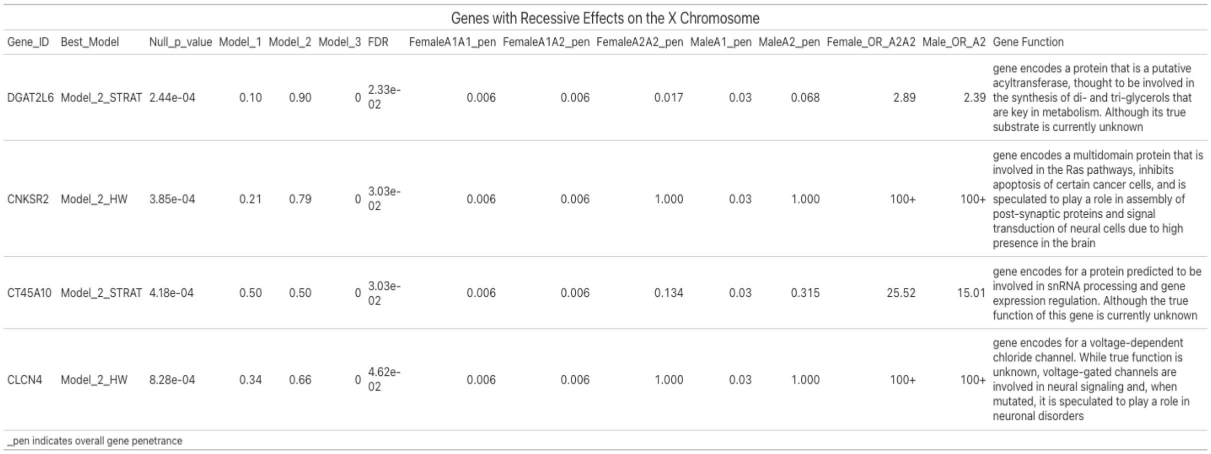
Genes with Additive Effects on the X Chromosome

Gene_ID	Best_Model	Null_p_value	Model_1	Model_2	Model_3	FDR	FemaleA1A1_pen	FemaleA1A2_pen	FemaleA2A2_pen	MaleA1_pen	MaleA2_pen	Female_OR_A1A2	Female_OR_A2A2	Male_OR_A2	Gene Function
ARHGEF9	Model_1_HW	5.12e-10	1.00	0.00	0	2.31e-07	0.006	0.688	1.000	0.03	1.000	100+	100+	100+	gene encodes a GTPase that aids in regulation of many genes involved in brain development and function. Defects in this gene result in epilepsy and other types of cognitive disabilities
MECP2	Model_1_HW	9.28e-10	1.00	0.00	0	2.31e-07	0.006	0.688	1.000	0.03	1.000	100+	100+	100+	gene encodes a transcriptional repressor protein that plays an essential role in mammalian development, specifically for embryonic development. A defective MECP2 gene is best known as the causes of the majority of cases of Rett Syndrome, a progressive neurologic development disorder in females
MAGEC1	Model_1_STRAT	3.89e-06	0.99	0.01	0	7.06e-04	0.006	0.082	0.390	0.03	0.635	14.63	100+	56.92	gene encodes for a protein that is a tumor-specific antigen responsible for immune system function, specifically in regulation of T lymphocytes
IQSEC2	Model_1_HW	2.49e-05	1.00	0.00	0	3.62e-03	0.006	0.399	0.975	0.03	0.995	100+	100+	100+	gene encodes a protein involved in the post-synaptic density of excitatory synapses. This encoded protein plays a crucial role in synaptic organization, with defects in this gene being involved in cognitive disabilities
RIPPLY1	Model_1_STRAT	9.34e-05	1.00	0.00	0	1.13e-02	0.006	0.688	1.000	0.03	1.000	100+	100+	100+	gene encodes a protein known to be a transcriptional repressor
PCDH19	Model_1_STRAT	1.82e-03	0.98	0.02	0	9.23e-02	0.006	0.491	0.993	0.03	0.999	100+	100+	100+	gene encodes a protein involved in calcium-dependent cell adhesion that is primarily expressed in the brain. Mutations on this gene have been implicated in infantile epilepsy

_pen indicates overall gene penetrance

Of the 15 genes that met FDR criteria, six were found to have an additive overall effect on ASD, corresponding to model 1 with near 100% posterior probability, suggesting that an additive model of disease fit the data dramatically better than either the recessive or homozygous lethal model. All FDR corrected p-values post-BH correction are shown above with all meeting minimum threshold criteria of <0.1, as well as unadjusted p-values produced after the intital QXL-TADA analysis. Penetrance for each of the of the additive genes is laid out by genotype (A1A1, A1A2, A2A2) for females, and for males (A1-, A2-). Each of those penetrances were subsequently converted to odds ratios (ORs) (*see methods section*) using the penetrance of A1A1 as the references group for females, and A1(-) as the reference group for males. Each gene ID is also marked with its corresponding function in vivo.

Table 2- Recessive Effect Genes Associated with ASD



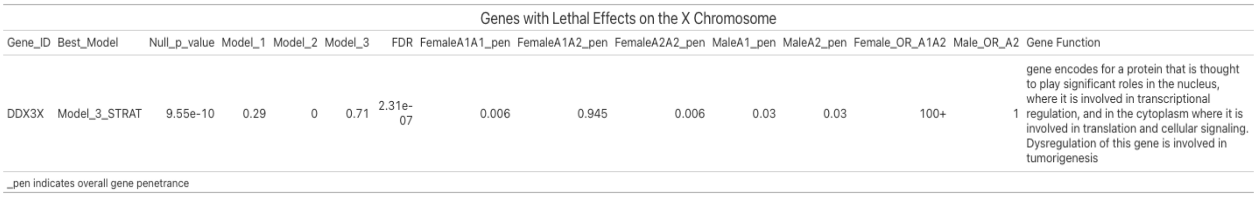
Genes with Recessive Effects on the X Chromosome

Gene_ID	Best_Model	Null_p_value	Model_1	Model_2	Model_3	FDR	FemaleA1A1_pen	FemaleA1A2_pen	FemaleA2A2_pen	MaleA1_pen	MaleA2_pen	Female_OR_A2A2	Male_OR_A2	Gene Function
DGAT2L6	Model_2_STRAT	2.44e-04	0.10	0.90	0	2.33e-02	0.006	0.006	0.017	0.03	0.068	2.89	2.39	gene encodes a protein that is a putative acyltransferase, thought to be involved in the synthesis of di- and tri-glycerols that are key in metabolism. Although its true substrate is currently unknown
CNKSR2	Model_2_HW	3.85e-04	0.21	0.79	0	3.03e-02	0.006	0.006	1.000	0.03	1.000	100+	100+	gene encodes a multidomain protein that is involved in the Ras pathways, inhibits apoptosis of certain cancer cells, and is speculated to play a role in assembly of post-synaptic proteins and signal transduction of neural cells due to high presence in the brain
CT45A10	Model_2_STRAT	4.18e-04	0.50	0.50	0	3.03e-02	0.006	0.006	0.134	0.03	0.315	25.52	15.01	gene encodes for a protein predicted to be involved in snRNA processing and gene expression regulation. Although the true function of this gene is currently unknown
CLCN4	Model_2_HW	8.28e-04	0.34	0.66	0	4.62e-02	0.006	0.006	1.000	0.03	1.000	100+	100+	gene encodes for a voltage-dependent chloride channel. While true function is unknown, voltage-gated channels are involved in neural signaling and, when mutated, it is speculated to play a role in neuronal disorders

_pen indicates overall gene penetrance

Of the 15 genes that meant FDR criteria, four of these genes were found to have an overall recessive effect related to ASD. Unlike the additive models that were all near 100% posterior probability, the recessive mode of inheritance was far less certain with posterior probabilities ranging from 90% certainty to around 50% certainty. All FDR corrected p-values post-BH correction are shown above with all meeting minimum threshold criteria of <0.1, as well as initial uncorrected p-values produced after the initial QXL-TADA analysis. The penetrance of each gene and corresponding ORs using A1A1 as reference for females, and A1(-) reference for males is also shown above. Subsequently, A1A2 females are absent since equal effect is seen on the heterozygotes compared to common allele (A1A1) homozygotes in a recessive model. Corresponding functions of the in vivo effect of the recessive genes is also given.

Table 3- Lethal Effect Genes associated with ASD



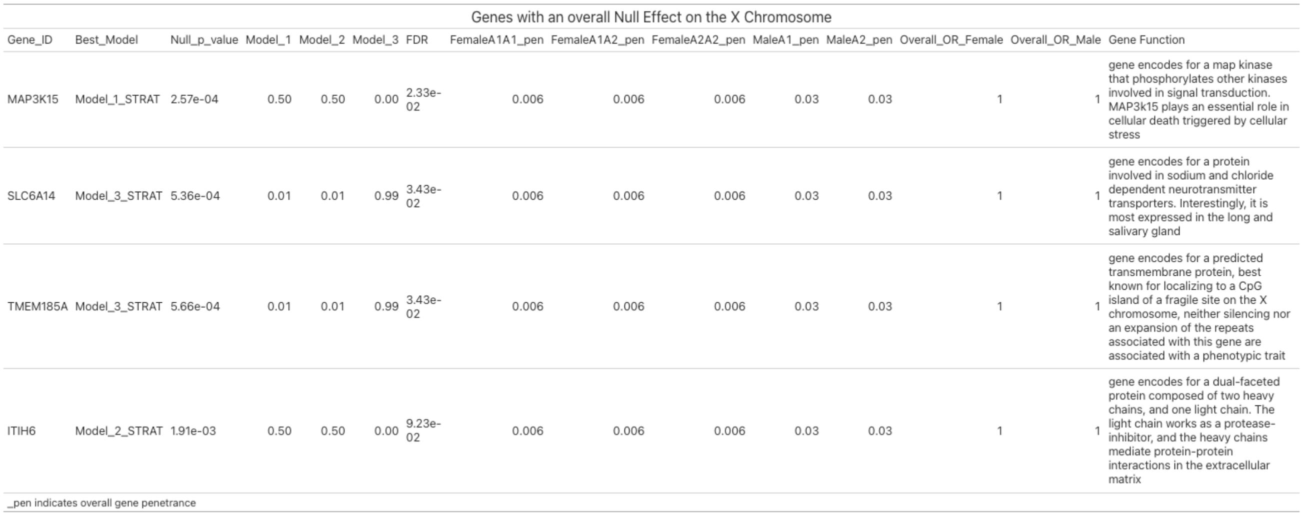
Genes with Lethal Effects on the X Chromosome

Gene_ID	Best_Model	Null_p_value	Model_1	Model_2	Model_3	FDR	FemaleA1A1_pen	FemaleA1A2_pen	FemaleA2A2_pen	MaleA1_pen	MaleA2_pen	Female_OR_A1A2	Male_OR_A2	Gene Function
DDX3X	Model_3_STRAT	9.55e-10	0.29	0	0.71	2.31e-07	0.006	0.945	0.006	0.03	0.03	100+	1	gene encodes for a protein that is thought to play significant roles in the nucleus, where it is involved in transcriptional regulation, and in the cytoplasm where it is involved in translation and cellular signaling. Dysregulation of this gene is involved in tumorigenesis

_pen indicates overall gene penetrance

Of the 15 genes that met FDR criteria, only one gene was found to have a true lethal effect and apparent heterozygous association with ASD in females. The posterior confidence of the lethal model was around 71%. The only FDR corrected p-value post-BH correction is shown above meeting minimum threshold criteria of <0.1, as well as the uncorrected p-value produced after the initial QXL-TADA analysis. The penetrance of DDX3X and the corresponding ORs are also referenced, again using A1A1 for females and A1(-) for males as a reference. Subsequently, the A2A2 homozygous recessive containing two effect alleles is left out since the overall OR is 1.00, indicating that A2A2 homozygotes are dead in this model. A2(-) genotype for males is left in for reference since we would not see males with A2 allele for a lethal gene like DDX3X. Corresponding function of the in vivo effect of the lethal gene is also given.

<u>Table 4- Null Effect Genes associated with ASD (Indicates one of the three models fit better than the null model, but there was no effect)</u>



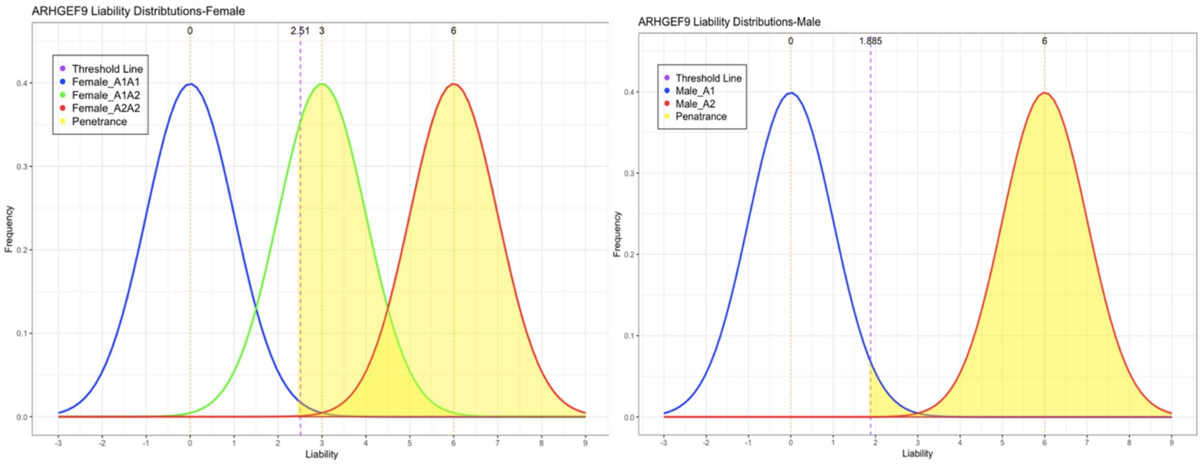
Genes with an overall Null Effect on the X Chromosome

Gene_ID	Best_Model	Null_p_value	Model_1	Model_2	Model_3	FDR	FemaleA1A1_pen	FemaleA1A2_pen	FemaleA2A2_pen	MaleA1_pen	MaleA2_pen	Overall_OR_Female	Overall_OR_Male	Gene Function
MAP3K15	Model_1_STRAT	2.57e-04	0.50	0.50	0.00	2.33e-02	0.006	0.006	0.006	0.03	0.03	1	1	gene encodes for a map kinase that phosphorylates other kinases involved in signal transduction. MAP3k15 plays an essential role in cellular death triggered by cellular stress
SLC6A14	Model_3_STRAT	5.36e-04	0.01	0.01	0.99	3.43e-02	0.006	0.006	0.006	0.03	0.03	1	1	gene encodes for a protein involved in sodium and chloride dependent neurotransmitter transporters. Interestingly, it is most expressed in the long and salivary gland
TMEM185A	Model_3_STRAT	5.66e-04	0.01	0.01	0.99	3.43e-02	0.006	0.006	0.006	0.03	0.03	1	1	gene encodes for a predicted transmembrane protein, best known for localizing to a CpG island of a fragile site on the X chromosome, neither silencing nor an expansion of the repeats associated with this gene are associated with a phenotypic trait
ITIH6	Model_2_STRAT	1.91e-03	0.50	0.50	0.00	9.23e-02	0.006	0.006	0.006	0.03	0.03	1	1	gene encodes for a dual-faceted protein composed of two heavy chains, and one light chain. The light chain works as a protease-inhibitor, and the heavy chains mediate protein-protein interactions in the extracellular matrix

_pen indicates overall gene penetrance

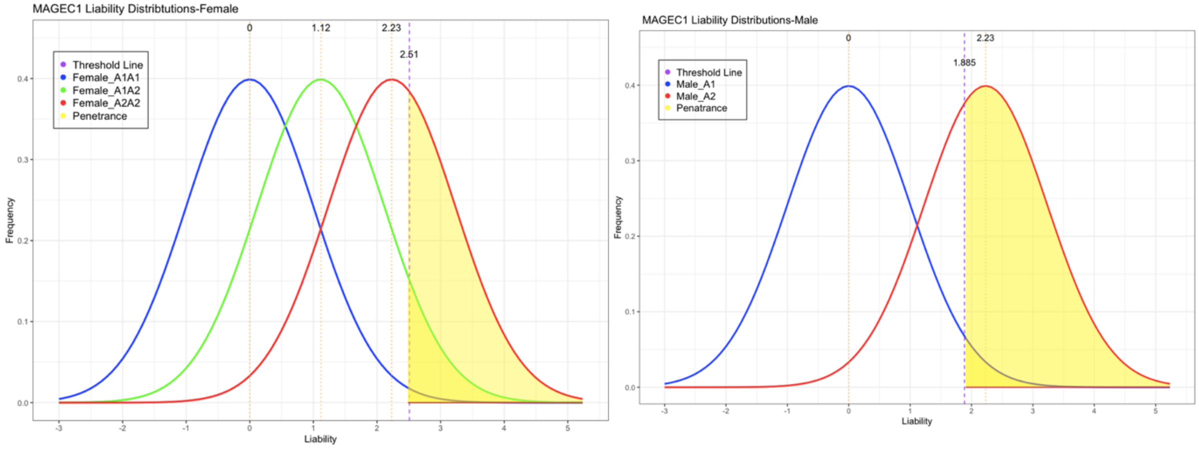
Of the 15 genes that met FDR criteria, four of these genes, although associated closely with one of our 3 models, had no effect on the incidence of ASD. Each of the 4 genes corresponded to different models with one gene each corresponding to models 1 and 2, and two genes corresponding to model 3. Both genes corresponding to model 3 had an extremely good fit with near 100% confidence, whereas the other two genes were a 50% split between model 1 and 2. All FDR corrected p-values post-BH correction are shown above with all meeting threshold criteria of <0.1, as well as initial null p-values produced after the initial QXL-TADA analysis. Each of these genes penetrances are laid out above; unlike other tables with varying odds ratios, this table lays out overall ORs for both male and female, as the null effects of each of the alleles resulted in ORs = 1.00 for each genotype. Corresponding functions for the genes that passed QC criteria but were ultimately null are also given.

Graphical Analysis and Penetrance Visualization- Additive Effect- Graph 1



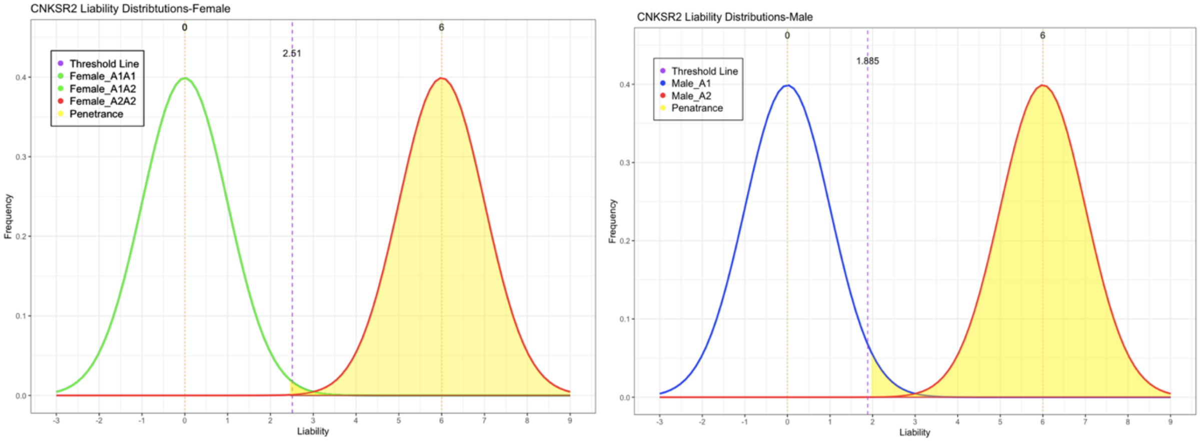
Graphical representation of the penetrance of additive genes can be visualized on normal distribution plots as described in the methods section. For additive effect variants for females, three curves are visualized indicating increasing penetrance from the common homozygote (A1A1) labeled in blue, heterozygote (A1A2) labeled in green, and rare homozygote (A2A2) labeled in red. In this example ARHGEF9 shows the effect of full penetrance (1.00) for the rare homozygote, indicated by the shaded area under the curve, and corroborated by the penetrance calculation in Table 1. Also visualized is the penetrance for the heterozygote (0.688), and the baseline penetrance of the common homozygote, that indicates a penetrance equal to the standard population prevalence (0.006). For the additive effect for males, there is an even starker contrast showing a fully penetrant rare hemizygote (A2(-), 1.000) labeled in red, and a baseline penetrance equal to the standard population prevalence for the common male hemizygote (A1(-), 0.030) labeled in blue. Thresholds are also shown as vertical asymptotes in light purple, indicating that individuals with liabilities above that threshold are those affected with ASD. Subsequently, liabilities calculated from initial QXL-TADA analysis are plotted as means, indicated by the orange asymptotes.

Graphical Analysis and Penetrance Visualization- Additive Effect- Graph 2



Above is another example of an additive effect gene, and corresponding penetrance for ASD. This example visualizes the effect of MAGEC1, a gene that is not fully penetrant in its rare homozygous form. For females, baseline penetrance for A1A1 remains the same (0.006), heterozygote penetrance for MAGEC1 is 0.082 (8.20%), and rare homozygote is 0.320 (32.0%). For males, baseline penetrance for the common hemizygote remains the same at 0.030 (3%), and rare hemizygote results in a penetrance of 0.635 (63.5%). Liabilities are shown on horizontal asymptote labeled in orange. Coloring and thresholds remain the same as previous graphs.

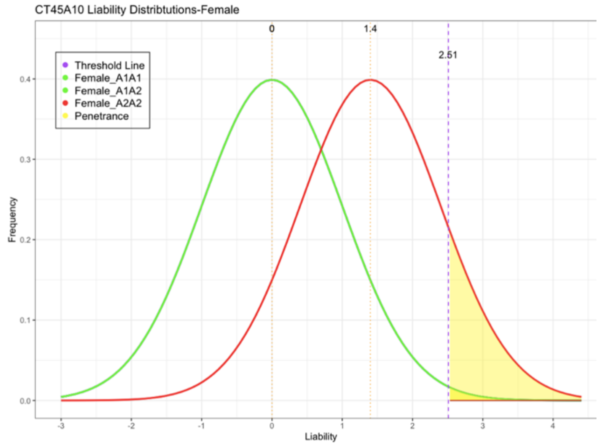
Graphical Analysis and Penetrance Visualization- Recessive Effect- Graph 1

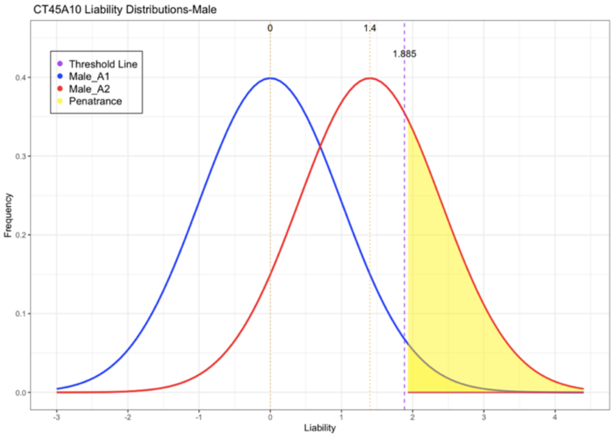

CNKSR2 Liability Distribtutions-Female

CNKSR2 Liability Distributions-Male

Similarly to the additive effect genes, recessive effect genes are also visualized via normal distribution plots, and penetrance is indicated by area under the curve, shaded in yellow. As our recessive modeled genes show in Table 2, penetrance of recessive genes show an effect only for the rare homozygote female (A2A2) and the rare hemizygote male (A2(-)) visualized by the graph for the CNKSR2 gene. In this case, both the common homozygote (A1A1) and the heterozygote (A1A2) genotypes of the CNKSR2, are shown in green as having a baseline liability of 0.006, where the rare homozygote (shown in red) confers near 100% penetrance. We see a similar phenomenon as related to the males hemizygotes, where the common hemizygote (A1(-)), modeled in blue, shows a baseline liability of 0.030, and the rare hemizygote (A2(-)), modeled in red, shows near 100% penetrance. Liabilities are shown on horizontal asymptote labeled in orange. Coloring and thresholds remain the same as previous graphs.

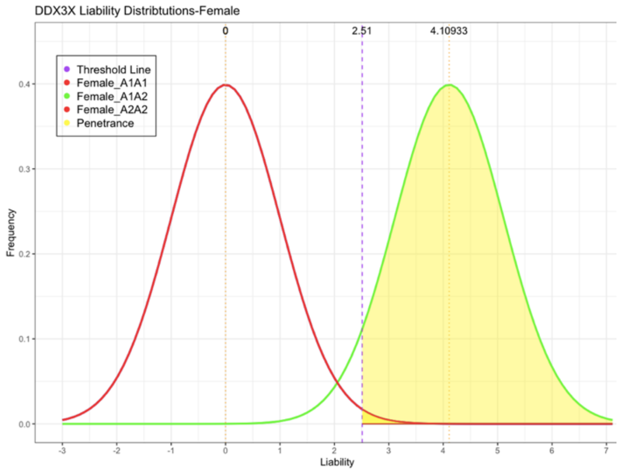
Graphical Analysis and Penetrance Visualization- Recessive Effect- Graph 2

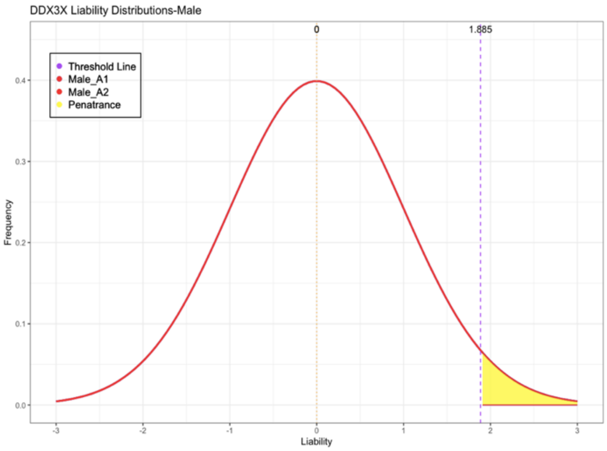




Above is another example of a gene following the recessive model, with shading corresponding to the penetrance of the gene CT45A10. Unlike the model for CNKSR2, this gene is not fully penetrant in its recessive form, conferring a rare homozygous penetrance of about 0.134 (13.4%) for females and 0.315 (31.5%) for males. Similarly, the female common homozygote (A1A1) and the heterozygote (A1A2), confer the same amount of penetrance as CNKSR2, which is the baseline liability, furthering evidence of the recessive effect of CT45A10. Liabilities are shown on horizontal asymptote labeled in orange. Coloring and thresholds remain the same as previous graphs.

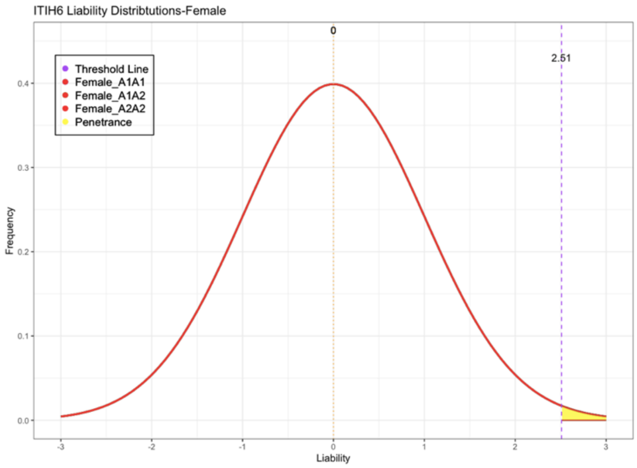
Graphical Analysis and Penetrance Visualization- Lethal Effect

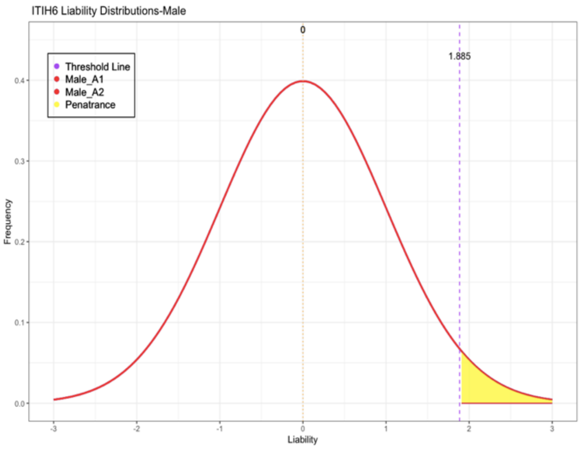




Lethal effect genes have a unique distribution and penetrance pattern, that do not follow a pattern seen by either the additive effect genes or the recessive effect genes. As our pattern of penetrance is shown in Table 3 and visualized above, the penetrance for the loan, lethal effect gene DDX3X shows only an effect for the female heterozygote (A1A2), and no effect for the common homozygote (A1A1), rare homozygote (A2A2), or either hemizygote (A1(-), A2(-)), subsequently modeled in red. Uniquely, the penetrance of the DDX3X heterozygote is the highest of any of the heterozygote effect genes modeled in the any of the graphs at 94.5% penetrance for ASD, modeled in green above. Liabilities are shown on horizontal asymptote labeled in orange. Coloring and thresholds remain the same as previous graphs.

Graphical Analysis and Penetrance Visualization- Null Effect





Null effect genes, shown in table 4, are genes that meant initial QC criteria (null p-value < 0.05, FDR <0.10) and were associated closely with model 1, model 2, or model 3, but upon penetrance calculation, the genes appeared to confer no liability to ASD. The graphs above, visualize this null effect, and no further consideration of the genes functions, nor liability for ASD, will be considered for further analysis.

DISCUSSION

Genetic Breakdown by Effect

This study analyzed more than 74,000 individuals from familial and case control data, including around 20,000 individuals with ASD. Upon analyzing all 726 genes with data from the X chromosome, and establishing strict QC criteria, we were able to implicate 11 genes that confer an overall risk for ASD at various thresholds of FDR of 0.10, 0.05, and 0.01, after BH correction. This identified genes that appear to have functions related to neural networks and transcriptional repression associated with cognitive and developmental delay often associated with ASD. Of the

11 genes that meant criteria for secondary analysis, it was found that 6 of these genes have functions that are associated with neurological function (ARHGEF9, MECP2, IQSEC2, PCDH19, CNKSR2, CLCN4), and 4 of the genes were involved in transcriptional repression (MECP2, RIPPLY1, CT45A10, DDX3X). It is worth noting that 9 genes in total were associated with either transcription repression or some cognitive process; however, 2 more genes were associated with other cellular processes. The gene DGAT2L6 is a metabolic process gene involved in the transfer of acyl groups in the synthesis of di- and tri-glycerol, known to be key in metabolism, although the true substrate (protein made from the gene) is currently unidentified. Another gene named MAGEC1 is involved in T lymphocyte regulation, encoding for a tumor specific antigen that often aids in preventing cells from becoming cancerous.

In the overall analysis, the genes that best fit "Model_1" most closely represent an additive model of penetrance/liability, meaning the effect of a heterozygote female (A1A2) should be directly in-between a common homozygote female (A1A1), and a rare homozygote female (A2A2). Effectively, this pattern was shown for 6 genes by likelihood ratio tests that best associated with Model_1 (p-value <0.002, FDR<0.1). Some of these genes have been implicated in single/candidate gene studies to be greatly associated with neurological development. For example, ARHGEF9 is a gene whose defectiveness is best associated with regulation of the neurotransmitter GABA and neuron excitability **(27)**. Best implicated in childhood epilepsy, this gene obviously is known to have some sort of neurological effect. In this analysis, ARHGEF9 was the most significant of our additive findings. Establishing a new connection between a rare variant of ARHGEF9 and an additive form of liability for ASD (p=5.12e-10, FDR < 0.01). Another gene worth mentioning is MECP2, which was shown in Table 1 as both simultaneously contributing to transcriptional repression, as well as the key gene that is mutated in Rett

Syndrome. In our study, MECP2 is implicated very strongly in an additive form of liability to ASD (p<9.28e-10, FDR <0.01), and is shown in our penetrance calculations (much like ARHGEF9) as being fully penetrant for ASD when carrying the rare allelic variant. This finding of its strong association with ASD is consistent with previous studies **(28, 29)** and is arguably one of the most well-known X-linked neurodevelopmental gene. Additionally, IQSEC2 (p = 2.49e-5, FDR < 0.01) and PCDH19 (p = 1.82e-3, FDR < 0.1) were also found to be significantly associated with ASD, per the additive model. It is known that IQSEC2 is a neurological gene in involved in excitatory synapse formation, as well as synapse organization. IQSEC2 has already been implicated in mild forms of intellectual disorders for those carrying a rare variant of the gene, and some small molecule therapies are said to be in the works to treat the defunct form of this gene as of 2023 **(30)**. PCDH19 is the final gene to be associated with both the additive model for ASD, and neurological function within the X chromosome. This gene is thought to be involved in the process of calcium-dependent cellular adhesion that is primarily expressed in the brain and has also been associated with the inhibitory neurotransmitter GABA. Neurologically, monogenetic epilepsy is a disorder closely associated with the rare allelic variant in PCDH19, and seizures related to the rare variant of PCDH19 are often triggered by photosensitivity **(31)**. This finding is peculiar in two ways. Firstly, this study's finding of a PCDH19 rare variant being associated with ASD sheds light on the fact that this gene may contribute to more neurological disorders than just epilepsy. Secondly, the fact that it has been shown that most seizures in those who have PCDH19 defects are often triggered by light sensitivity may shed light on a reason why a lot of those who display ASD are often sensitive to bright lights, which is one of the most common hallmarks of diagnosing ASD in the DSM-V **(32)**.

Genes that followed a recessive pattern of liability via likelihood ratio test, represented by "Model_2", were also implicated in the manifestation of ASD. Recessive pattern of ASD liability/penetrance is demonstrated by the A1A1 common female homozygote carrying the same liability as the A1A2 heterozygote, with the A2A2 rare homozygote carrying an increased liability for ASD. Four genes were implicated to follow a "Model_2" recessive pattern of liability for ASD corresponding to genes: DGAT2L6, CNKSR2, CT45A10, and CLCN4. Consistent with prior analysis, two of these genes that follow the recessive model are known to have significant effects on neural cells and pathways. CNKSR2 is a gene found to be involved in Ras pathway signal transduction and has been implicated to be involved in assembly and development of dendritic spines on primary neurons **(33)**. As of 2022, CNKSR2 has recently been described as a "causative gene" for X-linked syndromic mental retardation as well as X-linked intellectual disabilities resulting in cognitive delay, attention deficit, and early-onset seizures. Notably, our analysis pins down that CNKSR2 is wholly associated with ASD ($p = 3.85e-4$, $FDR<0.01$), which is mentioned in recent literature, but is often grouped in broadly with other neurological disorder criteria **(34)**. Another recessive model associated neurological gene implicated in conferring liability to ASD is CLCN4. CLCN4 is a gene that is involved in voltage-dependent chlorine channels in neural cells, although the exact function of how CLCN4 is involved remains partially unknown. Candidate gene studies of CLCN4 have implicated rare variants of this gene to be associated with neurodevelopmental delays, mental disorders, and intellectual disability **(35)**. Although known to be associated with cognitive delays, mutations and variants are known to be rare, and little research has been done on the underpinnings of this particular gene. This study gives a well associated link between the cognitive delays that are associated with the rare variant of CLCN4 and ASD ($p = 8.28e-4$, $FDR < 0.05$), as well as direct evidence of the likely recessive pattern of expression (likelihood = 66%). Two other genes were also shown to be associated with ASD following a recessive model of expression, although not known to be involved in neural pathways; these genes are DGAT2L6 and CT45A10. DGAT2L6 was the most significantly associated with ASD among the recessive models ($p = 2.44e-4$, $FDR < 0.05$), and is known to be a putative acyltransferase involved in the synthesis of di- and tri-glycerol in metabolism. Subsequent research on the ASD front is non-existent, but the rare variants of DGAT2L6 are known to be involved in metabolism for cancer cells and drive cellular immortalization **(36)**. Nevertheless, our analysis shows that some form of metabolic dysregulation is synonymous with ASD, but since the true substrate of DGAT2L6 is unknown, and no prior research has shown effect on neurological disorders, further research will need to be done to fully implicate this gene as causative for ASD. As this analysis shows, there is only a moderate increase in likelihood for ASD even in rare homozygous females (OR = 2.89), compared to other genes analyzed in this study that are most surely causative. Finally, the last gene implicated in the recessive model that confers increased likelihood for ASD is CT45A10. CT45A10 is a gene involved in snRNA processing and transcriptional regulation and is expressed in the brain. More in depth functions are currently unknown at this time, but I hypothesize, much like other genes in this study that are associated with ASD, it is a functional gene involved in building and modulating neural networks. It is worth noting with regards to the recessive effect model of CT45A10 that it was only a 50% likelihood upon initial analysis that CT45A10 followed a recessive liability pattern, subsequent analysis shown in Table 2 confirms that upon penetrance calculations, we do see CT45A10 follow a recessive penetrance pattern, and subsequently most associated with ASD via model 2 ($p = 4.18e-4$, $FDR < 0.05$).

Finally, there was only one gene that followed a lethal pattern of liability for ASD via likelihood ratio test represented by “Model_3” in our analysis. Lethal pattern of penetrance/liability is the most complicated due to the lethality of the rare homozygote females (A2A2) and the rare hemizygote males (A2(-)). In this case, the only genotypic variant that confers any liability/penetrance to ASD would be the female heterozygote, which is the case for the only gene that follows this model: DDX3X. DDX3X is a gene that is directly involved in transcriptional regulation, translation, and cellular signaling. Subsequent dysregulation of this gene is often involved in creation of tumors. DDX3X is the loan gene in our analysis that followed a lethal pattern of liability for ASD ($p = 9.55e-10$, $FDR < 0.01$) and is not present in males. While not being stated as a direct functional neurological gene, DDX3X is closely associated with neurological delays and extreme cognitive deficits. A study published in 2020 analyzing DDX3X found that rare variants of the DDX3X gene were associated with severe neurodevelopmental delays, such as most females going non-verbal by the age of 5-years-old. Interestingly, the study also states the rarity of DDX3X variants in males, perhaps alluding to further evidence of the lethal pattern of inheritance represented by the rare variants of this gene **(37)**. However, it is stated that deficits represented by other variants of this gene with regards to neurodevelopmental delays for males are not impossible. This could simply be because representation of other variants conferring neurological disorders other than ASD are possible in males, but in the context of this paper, no such evidence was found (see Table 3). Lethal effects of the penetrance pattern of DDX3X are best represented in the fact that the OR’s for both the rare homozygote females, and the rare hemizygote males are both 1; yet this is simply a statistical measure, as the true nature of individuals with this genotype do not exist in this study due to the lethal effects of the rare variant.

ASD Prevalence Differences by Sex

Prevalence differences of ASD between sexes were widely discussed at the beginning of this paper and have been further indicated through the differences in penetrances and odds ratios between sexes. Interestingly, throughout the analysis, and upon penetrance calculations, there is a strong indication that many rare genetic variants analyzed implicate a causative effect for ASD. In other words, genes such as ARHGEF9 and MECP2, and associated rare variants, show a near 100% penetrance for ASD, in both the rare homozygous female and rare hemizygous male. In this case, rare variants in these genotypes appear to confer ASD at the same rate, however the heterozygous females, on the additive scale, always have a lower penetrance than males. In terms of additive effect genes, one can assume that this decreased penetrance, perhaps given by the protective effect of the common allelic variant in the heterozygote females, contribute to some form of disparity between the rates males and females are affected with ASD. This could be explained by the fact that males do not have the genotypic ability to hold 2 alleles on their X chromosome and are subsequently at a higher risk. Thus, percentage-wise based on genotypic make up alone, the chance of holding a rare allele that confers full penetrance to ASD is higher. However, on the contrary, penetrance of the alleles only explains part of the disparity seen in odds of liability for ASD. Interestingly, males are not always at a higher risk for ASD when carrying rare variants. Surprisingly, in some cases, such as the rare variant for the gene MAGEC1, females who carry a rare homozygous genotype are at higher risk of ASD (OR = 100+) than their male counterparts (OR = 56.92), even though male penetrance for the rare genetic variant appears higher. Similar to the example in which a common variant can be protective, to some extent, against the penetrance of a rare variants that confer ASD; it appears it can also be the case that two rare variants, for rare female homozygotes, can lead to an even greater increased liability for ASD diagnosis. The recessive genes DGAT2L6 and CT45A10 also show examples of this disparity between the rare female homozygotes and rare male hemizygotes, further explaining this phenomenon. The debate between how these genotypes and associated penetrances of rare variants of X chromosome genes contribute to disparity between males and female with regards to prevalence of ASD in no way could explain all of the difference in diagnostic rate. However, this finding is worth noting since prior research on the autosomes did not find a substantial prevalence difference in rare variants of the autosomes **(2)** that are associated with ASD.

PUBLIC HEALTH IMPLICATIONS AND FUTURE DIRECTIONS

Future contributions to the field of human genetics, genetic epidemiology and biostatistics will no doubt take genetic analysis of human disease to new heights and expand upon the contribution laid out in this study. The future of genetic medicine and gene therapy to treat genetic deficits and disorders such as ASD is something that is on the horizon and will eventually aid in the cure of many incurable disease being studied today. Genetic association studies analyzing genetic defects are crucial in aiding the progression of genetic medicine that will bring numerous cures to society that were once thought to be impossible. The numerous resources geared towards supportive care for many genetic diseases and long term care facilities will ultimately be able to be re-allocated to getting individuals back on their feet and cured of aliments that have plagued many lives for far too long. Genetic epidemiology is a rapidly growing field, and we are just now starting to see the benefits of genetic studies pay off and make real world progress in contributing to decreased burden of disease, much in the way vaccinations revolutionized the way our society thinks of infectious disease. The rise in whole genome sequencing (WGS) studies will be able to give unprecedented access to the human genome that will no doubt rapidly increase knowledge about the level of complexity related to genetics, epigenetics, environmental exposures, and the contribution of variants in the human genome to disease. However, without studies like the one presented today, the field would remain stagnant and without a starting point for far more complex studies that will ultimately result in the decreased burden of genetic disease for years to come.

REFERENCES

1. Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., Awashti, S., Belliveau, R., Bettella, F., Buxbaum, J. D., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Christensen, J. H., Churchhouse, C., … Børglum, A. D. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, *51*(3), 431–444. https://doi.org/10.1038/s41588-019-0344-8
2. Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J. Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., Brand, H., … Buxbaum, J. D. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, *180*(3), 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036

3. Mordaunt, C. E., Jianu, J. M., Laufer, B. I., Zhu, Y., Hwang, H., Dunaway, K. W., Bakulski, K. M., Feinberg, J. I., Volk, H. E., Lyall, K., Croen, L. A., Newschaffer, C. J., Ozonoff, S., Hertz-Picciotto, I., Fallin, M. D., Schmidt, R. J., & LaSalle, J. M. (2020). Cord blood DNA methylome in newborns later diagnosed with autism spectrum disorder reflects early dysregulation of neurodevelopmental and X-linked genes. *Genome medicine*, *12*(1), 88. https://doi.org/10.1186/s13073-020-00785-8
4. Maenner, M. J., Shaw, K. A., Bakian, A. V., Bilder, D. A., Durkin, M. S., Esler, A., Furnier, S. M., Hallas, L., Hall-Lande, J., Hudson, A., Hughes, M. M., Patrick, M., Pierce, K., Poynter, J. N., Salinas, A., Shenouda, J., Vehorn, A., Warren, Z., Constantino, J. N., DiRienzo, M., … Cogswell, M. E. (2021). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002)*, *70*(11), 1–16. https://doi.org/10.15585/mmwr.ss7011a1
5. Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(6), 466–474. https://doi.org/10.1016/j.jaac.2017.03.013
6. Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E. J. H., Jones, R. M., Pickles, A., State, M. W., Taylor, J. L., & Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nature reviews. Disease primers*, *6*(1), 5. https://doi.org/10.1038/s41572-019-0138-4

7. Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: implications for explaining autism. *Science (New York, N.Y.)*, *310*(5749), 819–823. https://doi.org/10.1126/science.1115455

8. Napolitano, A., Schiavi, S., La Rosa, P., Rossi-Espagnet, M. C., Petrillo, S., Bottino, F., Tagliente, E., Longo, D., Lupi, E., Casula, L., Valeri, G., Piemonte, F., Trezza, V., & Vicari, S. (2022). Sex Differences in Autism Spectrum Disorder: Diagnostic, Neurobiological, and Behavioral Features. *Frontiers in psychiatry*, *13*, 889636. https://doi.org/10.3389/fpsyt.2022.889636

9. Sedgewick, F., Hill, V., Yates, R., Pickering, L., & Pellicano, E. (2016). Gender Differences in the Social Motivation and Friendship Experiences of Autistic and Non-autistic Adolescents. *Journal of autism and developmental disorders*, *46*(4), 1297–1306. https://doi.org/10.1007/s10803-015-2669-1

10. Juchniewicz, P., Piotrowska, E., Kloska, A., Podlacha, M., Mantej, J., Węgrzyn, G., Tukaj, S., & Jakóbkiewicz-Banecka, J. (2021). Dosage Compensation in Females with X-Linked Metabolic Disorders. *International journal of molecular sciences*, *22*(9), 4514. https://doi.org/10.3390/ijms22094514

11. Head, A. M., McGillivray, J. A., & Stokes, M. A. (2014). Gender differences in emotionality and sociability in children with autism spectrum disorders. *Molecular autism*, *5*(1), 19. https://doi.org/10.1186/2040-2392-5-19

12. Lai, M. C., Lombardo, M. V., Chakrabarti, B., Ruigrok, A. N., Bullmore, E. T., Suckling, J., Auyeung, B., Happé, F., Szatmari, P., Baron-Cohen, S., & MRC AIMS Consortium (2019). Neural self-representation in autistic women and association with 'compensatory

camouflaging'. *Autism : the international journal of research and practice*, *23*(5), 1210–1223. https://doi.org/10.1177/1362361318807159

13. Migeon B. R. (2020). X-linked diseases: susceptible females. *Genetics in medicine : official journal of the American College of Medical Genetics*, *22*(7), 1156–1174. https://doi.org/10.1038/s41436-020-0779-4
14. Leitão, E., Schröder, C., Parenti, I. *et al.* Systematic analysis and prediction of genes associated with monogenic disorders on human chromosome X. *Nat Commun* **13**, 6570 (2022). https://doi.org/10.1038/s41467-022-34264-y
15. Mallard, T. T., Liu, S., Seidlitz, J., Ma, Z., Moraczewski, D., Thomas, A., & Raznahan, A. (2021). X-chromosome influences on neuroanatomical variation in humans. *Nature neuroscience*, *24*(9), 1216–1224. https://doi.org/10.1038/s41593-021-00890-w
16. Samocha KE, Kosmicki JA, Karczewski KJ, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv; 2017. DOI: 10.1101/148353.
17. Zhang, X., Below, P., Naj, A., Kunkle, B., Martin, E., & Bush, W. S. (2023). Predicting genetically regulated gene expression on the X chromosome. *bioRxiv : the preprint server for biology*, 2023.06.06.543877. https://doi.org/10.1101/2023.06.06.543877
18. Laumonnier, F., Cuthbert, P. C., & Grant, S. G. (2007). The role of neuronal complexes in human X-linked brain diseases. *American journal of human genetics*, *80*(2), 205–220. https://doi.org/10.1086/511441
19. Pfeiffer, B. E., & Huber, K. M. (2009). The state of synapses in fragile X syndrome. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, *15*(5), 549–567. https://doi.org/10.1177/1073858409333075

20. Bartholomay, K. L., Lee, C. H., Bruno, J. L., Lightbody, A. A., & Reiss, A. L. (2019). Closing the Gender Gap in Fragile X Syndrome: Review on Females with FXS and Preliminary Research Findings. *Brain sciences*, *9*(1), 11. https://doi.org/10.3390/brainsci9010011

21. Na, E. S., & Monteggia, L. M. (2011). The role of MeCP2 in CNS development and function. *Hormones and behavior*, *59*(3), 364–368. https://doi.org/10.1016/j.yhbeh.2010.05.014

22. Petriti, U., Dudman, D. C., Scosyrev, E., & Lopez-Leon, S. (2023). Global prevalence of Rett syndrome: systematic review and meta-analysis. *Systematic reviews*, *12*(1), 5. https://doi.org/10.1186/s13643-023-02169-6

23. Jiang, Z., Sullivan, P. F., Li, T., Zhao, B., Wang, X., Luo, T., Huang, S., Guan, P. Y., Chen, J., Yang, Y., Stein, J. L., Li, Y., Liu, D., Sun, L., & Zhu, H. (2023). The pivotal role of the X-chromosome in the genetic architecture of the human brain. *medRxiv : the preprint server for health sciences*, 2023.08.30.23294848. https://doi.org/10.1101/2023.08.30.23294848

24. Fu, J. M., Satterstrom, F. K., Peng, M., Brand, H., Collins, R. L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S. P., Stevens, C. R., Cusick, C., Babadi, M., Banks, E., Collins, B., Dodge, S., Gabriel, S. B., Gauthier, L., Lee, S. K., Liang, L., … Talkowski, M. E. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nature genetics, 54(9), 1320–1331. https://doi.org/10.1038/s41588-022-01104-0

25. McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome biology, 17(1), 122. https://doi.org/10.1186/s13059-016-0974-4
26. Morton, N. E., & MacLean, C. J. (1974). Analysis of family resemblance. 3. Complex segregation of quantitative traits. American journal of human genetics, 26(4), 489–503.
27. Yang, H., Liao, H., Gan, S., Xiao, T., & Wu, L. (2022). ARHGEF9 gene variant leads to developmental and epileptic encephalopathy: Genotypic phenotype analysis and treatment exploration. Molecular genetics & genomic medicine, 10(7), e1967. https://doi.org/10.1002/mgg3.1967
28. Wen, Z., Cheng, T. L., Li, G. Z., Sun, S. B., Yu, S. Y., Zhang, Y., Du, Y. S., & Qiu, Z. (2017). Identification of autism-related MECP2 mutations by whole-exome sequencing and functional validation. Molecular autism, 8, 43. https://doi.org/10.1186/s13229-017-0157-5
29. Nagarajan, R. P., Hogart, A. R., Gwye, Y., Martin, M. R., & LaSalle, J. M. (2006). Reduced MeCP2 expression is frequent in autism frontal cortex and correlates with aberrant MECP2 promoter methylation. Epigenetics, 1(4), e1–e11. https://doi.org/10.4161/epi.1.4.3514
30. Levy, N. S., Borisov, V., Lache, O., & Levy, A. P. (2023). Molecular Insights into IQSEC2 Disease. International journal of molecular sciences, 24(5), 4984. https://doi.org/10.3390/ijms24054984
31. Moncayo, J. A., Ayala, I. N., Argudo, J. M., Aguirre, A. S., Parwani, J., Pachano, A., Ojeda, D., Cordova, S., Mora, M. G., Tapia, C. M., & Ortiz, J. F. (2022). Understanding

Protein Protocadherin-19 (PCDH19) Syndrome: A Literature Review of the Pathophysiology. Cureus, 14(6), e25808. https://doi.org/10.7759/cureus.25808

32. American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders. Text revision.
33. Ito, H., & Nagata, K. I. (2022). Functions of CNKSR2 and Its Association with Neurodevelopmental Disorders. Cells, 11(2), 303. https://doi.org/10.3390/cells11020303
34. Higa, L.A., Wardley, J., Wardley, C. et al. CNKSR2-related neurodevelopmental and epilepsy disorder: a cohort of 13 new families and literature review indicating a predominance of loss of function pathogenic variants. BMC Med Genomics 14, 186 (2021). https://doi.org/10.1186/s12920-021-01033-7
35. Palmer, E. E., Nguyen, M. H., Forwood, C., & et al. (2021, December 16). CLCN4-related neurodevelopmental disorder. In M. P. Adam, J. Feldman, G. M. Mirzaa, & et al. (Eds.), GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK575836/
36. Graber, M., Barta, H., Wood, R., Pappula, A., Vo, M., Petreaca, R. C., & Escorcia, W. (2021). Comprehensive Genetic Analysis of DGAT2 Mutations and Gene Expression Patterns in Human Cancers. Biology, 10(8), 714. https://doi.org/10.3390/biology10080714
37. Johnson-Kerner, B., Snijders Blok, L., Suit, L., & et al. (2020, August 27). DDX3X-related neurodevelopmental disorder. In M. P. Adam, J. Feldman, G. M. Mirzaa, & et al. (Eds.), GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK561282/