Endophenotype-Wide Association Study Reveals Genetic Substrates of Core Symptom Domains and Neurocognitive Function in Autism.

In-Hee Lee
Boston Children's Hospital
Ekaterina Koelliker
Boston Children's Hospital https://orcid.org/0000-0002-3310-2543
Sek Won Kong (sekwon.kong@childrens.harvard.edu)
Boston Children's Hospital https://orcid.org/0000-0003-4877-7567

Research

Keywords: Autism, endophenotype, core symptoms, cognitive function, ADOS, ADI-R

Posted Date: October 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-948337/v1

Title: Endophenotype-wide association study reveals genetic substrates of core symptom domains and neurocognitive function in autism

Authors:

In-Hee Lee, Ph.D.,[1] Ekaterina Koelliker,[2] Sek Won Kong, M.D.[1,3]

Affiliations:

1. Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115, USA
2. Psychology Department, Colby College, Waterville, ME 04901, USA
3. Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

Correspondence:

Sek Won Kong, M.D.

sekwon.kong@childrens.harvard.edu

40 Park Drive, LM5528.4

Boston, MA 02115, USA

Phone: 1-617-919-2689

Keywords:

Autism, endophenotype, core symptoms, cognitive function, ADOS, ADI-R

Abstract

Background

Autism is a neurodevelopmental disorder largely attributable to rare and common genetic variants. Additionally, environmental factors such as maternal immune activation and air pollution exposure can also increase the risk of autism. Genetic heterogeneity of autism has been well-recognized from gene discovery efforts over the past decade; however, genetic substrates of endophenotypes that constitute phenotypic heterogeneity are not known yet.

Methods

Whole-genome sequencing (WGS) data and a set of phenotype scores that represent neurocognitive development and the severity of core symptoms of autism were collected from the iHART and MSSNG databases and the phenotype database of Autism Speaks. Endophenotype-wide association analysis was performed with genome-wide genotype and 29 phenotype scores.

Results

One or more genetic loci were associated with each of phenotype scores at a genome-wide significance threshold (P=5×10^{-8}) except for a total score of the Social Responsiveness Scale-2. An intergenic locus on chromosome 15q26.1 was significant for three core symptom domain scores of ADOS Module 1 while each phenotype score was associated with a unique set of genetic loci. The Repetitive Behaviors Scale total score was associated with the largest number of loci (N=132) including the loci that overlapped with the genes involved in brain development and neurodegenerative disorders. Among the significant genotype-endophenotype associations, verbal intelligence and the *OSTN* gene was notable. The secretory peptide osteocrin—encoded by *OSTN*—is implicated in activity dependent dendritic growth in human and has potential for a biomarker of autism and an endophenotype marker for verbal intelligence.

Limitations

Validation of our findings in another cohort is required. Several associations involving the ADI-R and ADOS scores may indicate inherited allelic differences between affected and unaffected individuals since unaffected siblings were included in our analysis.

Conclusions

Our results suggest that autism candidate genes discovered by case-control GWAS may include trait-associated genes for core symptoms.

Background

Autism spectrum disorder (ASD) is a genetic neurodevelopmental disorder characterized by deficits in verbal communication and social interactions that co-occur with restricted, repetitive behaviors (RRBs). Impairments must be present in three core domains (communication, social, and behavior) in order for a diagnosis of ASD to be made.(1) The Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) are widely regarded as the "gold standard" for ASD diagnosis as they represent criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM).(2) Amongst individuals with ASD, phenotypic heterogeneity in adaptative functioning, cognitive development, and neurological deficits such as epilepsy, hydrocephalus, and sleep disorders is immense.(3) In the realm of cognitive functioning, intellectual disability (IQ ≤ 70) affects 33% of individuals with ASD.(4) As a result, the assessment of cognitive and adaptive abilities extending beyond the traditional triad of symptoms is useful for assessing ASD severity.(5) Moreover, accurate evaluation of such skills is crucial to understand the phenotypic heterogeneity as well as treatment strategy and outcomes.(5)

Genetic factors are the leading causes of ASD. A meta-analysis summarizing several decades of twin studies estimated that the heritability of ASD ranges from 0.64-0.91 as demonstrated by the discrepancy in concordance rates for monozygotic and dizygotic twins with ASD of unknown cause.(6) The vast inherited component of ASD is supported by familial clustering of cases(7) and higher concordance rates in siblings of individuals with autism (2-8%) in comparison to the general population.(8) As such, decades of gene discovery efforts using genotyping microarray and next-generation sequencing uncovered common and rare genetic variants that are enriched in ASD compared to neurotypical children. Several copy number variants (CNVs) have previously been associated with ASD. Together, genetic factors can be identified in no more than 20% of cases. Nonetheless, individuals with shared genetic risk factors do not present similar phenotypic profiles in the three core symptom domains.(9)

Diagnostic validity of ASD is well established (10) while genetic and phenotypic heterogeneity are evident.(3) Research Domain Criteria (RDoC) were established by the National Institute of Mental Health (NIMH) to create a framework for research on pathophysiology, especially for genomics and neuroscience, which will ultimately inform classification schemes.(11) The idea was to introduce a parallel categorization system to DSM-5, which describes validated dimensions of functioning relevant to mental health that can be linked to underlying biological systems. To this end, endophenotypes (EPs) enable researchers to narrow the gap between mental disorders and their genetic underpinnings. The commonly proposed models of EPs were reviewed by Kendler and Neale: the liability-index (or "risk-indicator") model and the mediational model.(12) The former mechanism postulates that risk for dichotomous mental disorders and continuous EPs are correlated with a common set of genes. On the other hand, the latter illustrates a causal pathway in which genetic variants influence EPs, leading to the corresponding mental disorder. Although Kendler and Neale noted the stronger and more falsifiable nature of the mediational model, EPs are explained most accurately with a bivariate or multivariate paradigm. In fact, several EPs of a disorder such as cognitive abnormalities and antisocial behavior in schizophrenia can be accounted by distinct components of genetic risk.(13)

Here we aimed to find genetic loci associated with EPs assessed by diverse instruments for various aspects of ASD and cognitive systems. We collected phenotype measures in diverse domains of neurocognitive function evaluated by standard instruments and tests—ADI-R, ADOS, Repetitive Behavior Scale, Revised (RBS),(14) Social Responsiveness Scale version 2 (SRS),(15) Peabody Picture Vocabulary Test III (PPVT),(16) Raven's Progressive Colored Matrices (RPCM),(17) Stanford-Binet Intelligence Scale, 5th edition (SB-5),(18) Vineland Adaptive Behavior Scale (VABS),(19) and head circumference (HC)— that are essential to evaluate positive and negative valences of ASD, as well as related cognitive systems and social processes in the context of RDoC framework. With detailed phenotype data and whole genome sequencing (WGS), we employed a genome-wide association (GWA) analysis framework to discover common genetic variants that are associated with the core symptoms of ASD. We found associations between neurocognitive features of ASD and several variants that have previously been described in the context of psychological disorders, supporting their likely contribution to the genetic underpinnings of ASD.

Method

Participants

Family-based data was collected from all individuals who participated in the Autism Genetic Resource Exchange (AGRE) Consortium, which compiles the WGS and phenotype data of families containing at least one individual with ASD diagnosed by the ADI-R and ADOS.(20) Although both instruments assess the three domains of ASD, they differ in format; the ADI-R is a structured caregiver interview that is shorter,(21) while the ADOS involves observation of the examinee in a series of standardized scenarios.(22) The ADI-R was utilized to characterize individuals in the sample as Autism, Not Quite Autism (NQA), Broad Spectrum, or Not Met.(21) In accordance with previous methods, we classified individuals as "case" if they fell under the Autism or NQA categories while "unaffected" individuals were those who were categorized as Broad Spectrum or Not Met by the AGRE. In addition to ASD-specific diagnostic tests, participants were given an opportunity to complete additional phenotype evaluations and these scores were utilized in GWA studies (GWASs).

Our AGRE dataset consisted of 11,961 individuals with demographic and phenotypic information, including 3,833 individuals with WGS data available. WGS data was collected through MSSNG and the Hartwell Autism Research and Technology Initiative (iHART) consortiums. MSSNG, a joint effort of Autism Speaks, University of Toronto, SickKids Hospital, and Google, is the largest collection of readily available WGS data for ASD researchers.(23) In its first phase of collection, MSSNG aimed to incorporate the phenotype scores and WGS data from individuals who were primarily part of the AGRE.(23) iHART is distinct in that its collection of WGS data from AGRE individuals focuses on multiplex families.(24) Both repositories have allowed for the successful identification of novel ASD-risk genes, which furthers our progress in developing interventions for the disorder. A summary of the demographic data for the entire AGRE dataset as well as for individuals with WGS data can be accessed in **Additional File 1**.

Phenotype scores

For our analysis, we analyzed 29 scores from nine phenotypic instruments compiled in the AGRE dataset: ADI-R, ADOS, RBS, SRS, PPVT, RPCM, SB-5, VABS, and HC. Each instrument covers one or more core symptom domains of ASD or neurocognitive development by age to access the severity of a participant's impairment. ADI-R, ADOS, and SRS have components to estimate difficulties in social interaction. RRBs are scored in the ADI-R, ADOS, and RBS while deficits in verbal and nonverbal communication are mostly measured by the ADI-R and ADOS. General neurocognitive development is estimated by RPCM, PPVT, SB-5, and VABS. **Additional File 2** summarizes the instruments and scores used in our study. The number of individuals with scores for each phenotype measure (either in the entire AGRE dataset or with WGS data available) varied because of the differences in compliance and completion rates across phenotypic instruments. Among the anthropometric measurements, we incorporated HC that is well-studied in the context of ASD and associated genetic conditions.(25)

Phenotype assessment for core symptoms of ASD

The ADI-R is a standardized, semi-structured interview administered by an experienced rater to a caregiver of participants suspected of having ASD. Effective for differentiating ASD from similar developmental disorders, the ADI-R is concerned with the participant's development, social functioning, language acquisition, and RRBs. In our study, we used the 4 corresponding domain scores– Social, Verbal and Nonverbal Communication, and Behavior.

Similarly, the ADOS is a standardized diagnostic test for ASD commonly used as a screening tool by school systems and clinicians. AGRE participants were administered ADOS Module 1, 2, or 3 at the discretion of a clinical psychologist according to their expressive language level. Through standardized scenarios, the test measures impairments in the domains of Social, Communication, Social-

Communication, Stereotyped Behaviors and Restricted Interests, and Play (Module 1 only). We used all of domain total scores available from each module (5 for Module 1, 4 for Modules 2 and 3) and the 3 total scores for each module, resulting in a total of 16 phenotype scores.

The RBS is a caregiver-informant questionnaire that quantifies various forms of RRBs that are characteristic of ASD.(14) Participants are evaluated on six subscales: stereotyped behavior, self-injurious behavior, compulsive behavior, ritualistic behavior, sameness behavior, and restricted behavior. The RBS Overall Score, which combines the five subscale (i.e., Ritualistic/sameness, Self-injurious, Stereotypic, Compulsive, and Restricted) scores, provides a measure of RRB severity and was chosen for our analysis. To encompass the distinct social domain of ASD, we also incorporated the SRS T-Score Total in our study. The SRS is a widely accepted measure of social impairment in the realms of social awareness, social cognition, social communication, social motivation, and mannerisms.

Instruments for assessing neurocognitive development

The SB-5 quantifies the cognitive abilities and intelligence of clinical and nonclinical populations.(18) The total scores from these two realms are combined to yield the full-scale IQ (FSIQ) score, which is used in addition to verbal IQ (VIQ) and nonverbal IQ (NVIQ) scores in the present study. All three of these scores are age-normed (mean 100, standard deviation (SD) 15). To provide additional information about each participant's neurocognitive development and encompass receptive vocabulary, we incorporated the PPVT standard score (mean 100, SD 15). The PPVT is an individually administered assessment of receptive lexical knowledge.(16) Of the three different versions recorded for the AGRE cohort, we chose 'Version 3' since it was used for most individuals with a reported PPVT score (1,681 out of 2,239).

Consisting of a series of tasks in which participants are required to identify missing elements of matrix patterns, the RCPM is a measurement of nonverbal intelligence.(26) The assessment serves as a paramount measurement of nonverbal processing, fluid intelligence, and spatial reasoning.(27) We utilized raw total scores from the RCPM in our analyses. The VABS is a semi-structured caregiver interview examining a participant's adaptive behavior and living skills.(28) An individual's level of functioning within the domains of communication, daily living skills, socialization, and motor skills are evaluated and used to derive the composite standard score– an age-normalized score (mean 100, SD 15) used for the purposes of the current investigation.

Genotype data

Merged variant call files were downloaded from the MSSNG (version db6, N=9,621) and iHART (version v01, N=2,308) project sites. After selecting individuals with available phenotype scores and filtering genotype data (bi-allelic variants of 0% genotype missing rate, variant allele frequencies between 5% and 95%, Hardy-Weinberg equilibrium), a total of 5,313,961 variants (4,983,916 single nucleotide variants (SNVs) and 330,045 indels) on autosomes across 3,833 individuals were available to be tested for association with phenotypic scores. The top 10 principal components (PCs) calculated from the 3,833 individuals were used as covariates to control for global ancestral backgrounds. For each phenotypic score, we used genetic variants with allele frequencies between 5% and 95% among individuals with the tested phenotypic score since the number of available individuals varied by phenotype score (**Additional File 3**). Therefore, the number of tested variants was less than 5,313,961 and varied across tests. For each phenotype score, we performed a GWA using PLINK (version v2.00a3LM downloaded from https://www.cog-genomics.org/plink/2.0/). Participant age at each test performed, gender, and top 10 PCs were used as covariates. A genome-wide significance *P*-value threshold of 5×10^{-8} was applied to select significant genomic variants for each analysis.(29) The summary statistic files from PLINK were used as an input to Functional Mapping and Annotation (FUMA, available at https://fuma.ctglab.nl/) for functional annotation and regional plots.(30)

To calculate the proportion of variance in phenotype score that was explained by genotype and the other covariates, we performed variance component analysis (VCA). The age at completion of each test, gender, genetic ancestry, and polygenic risk score (PRS) for ASD were utilized as covariates. PRS for ASD was calculated using the risk alleles at *P*-value <0.1 as reported by Grove *et al.*(31) using

PLINK. To make VCA computationally feasible, the numeric covariates (such as age at test and PRS) were discretized: 4 levels for age (7 years old or less; 7 to 9 years old; 9 to 12 years old; 12 years old or more) and decile ranks for PRS. For head circumference measurement, we grouped ages differently (7 years old or less; 7 to 12 years old; 12 to 18 years old; 18 years old or more) due to wide range of values (from 1.7 to 60.2 years old). The statistical language R (version 4.0.5) and the R library 'VCA' (version 1.4.3) were used for the analysis.

Results

Phenotype scores

The diagnostic and neurocognitive measurements used in the current study and the number of available participants for each measurement are listed in **Additional File 3**. Since the number of available phenotype scores varied across individuals, we used all participants for each phenotype score instead of selecting a subgroup (N=509) with all phenotype scores. Thus, each association test included a different number of individuals. For instance, the ADI-R social domain score was available for 3,746 individuals (includes 3,386 probands and 358 unaffected siblings) while the SB-5 FSIQ score was available for 833 individuals (includes 681 probands and 146 unaffected siblings). Between individuals with available WGS data and all individuals in the AGRE cohort, differences for all phenotype scores except for ADOS Module 3 Behavior Total, ADI-R Verbal Communication Total, RBS Overall Score, and SRS Total T-score were not significant at the threshold of P <0.01 (Wilcoxon test), confirming that the group with WGS data is an unbiased subset within the AGRE dataset. We also compared the distribution of scores in the AGRE cohort with published results. This comparison allowed us to check whether our cohort displayed any bias in terms of severity of ASD and neurocognitive traits. All scores were comparable with published baseline scores for individuals with ASD (**Additional File 3**).

Genome-wide Association Analysis of Core Symptoms of Autism

Participants were given an opportunity to complete the eight phenotype evaluations: ADI-R, ADOS, SRS, RBS, PPVT, RPCM, SB-5, and VABS, but compliance rates varied based on the test. For each phenotype measure, one or more domains were chosen to assess the severity of impairment. We performed GWA analysis for phenotype scores and HC measurement (N=29) using PLINK. A total of 681 variants were associated with 16 scores at the threshold of $P < 5\times10^{-8}$ (**Figure 1**). Each phenotype score was associated with a median of 1.5 variants (ranging from 1 to 531). For each test, PLINK output was uploaded to the FUMA server to identify risk loci from independent significant variants and annotation, resulting in 174 genomic risk loci (**Additional File 4** for detailed list and **Additional File 5** for regional plots). Of note, we found 68 loci where rare homozygous variants were observed in a small fraction of study cohort (<1% of entire tested individuals) with extreme phenotype scores, for which rare variants transmitted from both parents drove the significant associations. Except for the SRS, all phenotype tests measuring neurocognitive function were associated with one or more genetic loci. The leading variants for four phenotypes— ADI-R Verbal Communication Total and Nonverbal Communication Total scores, RBS Overall Score, and HC—passed the *P*-value threshold of 1.72×10^{-9} ($= 5\times10^{-8}/29$) adjusted for multiple concurrent hypothesis testing with 29 scores.

All four of the calculated total scores on the ADI-R (Social, Nonverbal Communication, Verbal Communication, and Behavior) that comprise the characteristic deficits of ASD were associated with one or more genetic loci (**Table 1** and **Additional File 4**). Behavior and Verbal Communication total scores were associated with intronic SNVs in the *HECW1* and *CDYL* genes, respectively. For an intronic variant in the *CDYL* gene (rs11754469, $P = 7.15\times10^{-10}$), both male and female carriers of the CC genotype displayed significantly decreased scores compared to individuals with TT and TC genotype groups (**Figure 2A**). The locus associated with the social domain of the ADI-R was found in 350kbps downstream of the *KLF6* gene.

The ADI-R Nonverbal Communication Total was associated with two noteworthy loci—intronic regions of *PCLO* and *SEMA3E*— at chromosome 7q21.11 (**Figure 2B**) for which structural variations such as microdeletions have been reported in ASD as well as other developmental delays. Presynaptic cytomatrix protein piccolo (*PCLO*) plays a role in monoamine neurotransmission(32) and presynaptic terminal enrichment.(33) Notably, *PCLO* is a candidate gene for schizophrenia(34) and ASD.(35) While previous findings related to the *SEMA3E* gene have not been specific to ASD, chromosome 7q21.11 microdeletions including this gene are described in patients with CHARGE syndrome (MIM# 214800) for which behavioral phenotype of autism are frequently reported.(36) Semaphorins and plexins are ligand-receptor pairs that regulate axon growth, and semaphorin 3E (*SEMA3E*) acts as both repellent and attractant depending on the presence of Neurophiln-1.(37)

For ADOS, distinct domain scores in Modules 1 and 3 were significantly associated with multiple loci while no loci were associated with Module 2 domain scores. Social and Social-Communication Total scores on Module 1 were associated with two and one loci, respectively, and the overall score was associated with three loci. Of note, a locus spanning the *HOX10* and *HOX11* genes was associated with the Module 1 Total score (**Figure 2C**). For ADOS Module 3, the domain total scores for behavior and the *PTPRD* gene represented the most significant association of the study.

The greatest number of significant loci (N=132) was found for RBS Overall Score, including several loci that overlapped with intronic or coding exons of genes involved in brain development (*MACROD2, PLCB1, TRAPPC9,* and *ZFHX3*), macrocephaly (*RIN2*), Parkinson's disease (*PANK2*), Alzheimer's disease (*GPAM*), immune disorders (*PEBP4*), cardiovascular disorders (*EDIL3, FMNL2, SCN5A,* and *TEC*), metabolic disorders (*NBAS* and *OSTN*), and other neurological disorders (*HDAC9*, *MICAL2*, *TEC*, and *TIAM1*) (**Additional File 4**). The two intronic loci within *MACROD2* are displayed in **Figure 3A**. Moreover, 8 loci were mapped to 94 genes (within 10kb of significant variants) that are enriched with candidate genes for ASD or schizophrenia from gene-set enrichment analysis with the GWAS catalog (Fisher's exact test, adjusted P=9.75×10^{-16}). Among the genes associated with significant loci for RBS, many are previously-reported candidate genes for ASD (e.g., *MACROD2, MEIS2, PARD3B, PEBP4, PLCB1, PTPRT,* and *TRAPPC9*) (**Additional File 4**).(38)

The *MACROD2* gene, which encodes mono-ADP-ribosylhydrolases, is evolutionarily conserved across mammalian species. *MACROD2* has multiple biological functions and is highly expressed in several regions of the developing and adult brain.(39) Independent GWASs for ASD(31),’(40) and traits resembling ASD in the general population(41) discovered risk alleles in the *MACROD2* gene. Rare copy number variation harboring *MACROD2* was found in patients with Attention Deficit Hyperactivity Disorder (ADHD).(42) Further investigations uncovered exon 5 deletion of this gene in patients with Kabuki syndrome, which is characterized by growth delays and intellectual disability.(43) Thus, *MACROD2* is a strong candidate gene for neurodevelopmental disorders, while its molecular function in the developing brain remains poorly understood. We also found that genes directly overlapping with significant variants were enriched with calcium ion binding functionality (Fisher's exact test, adjusted P = 0.030).

Genetic Substrates for Cognitive Systems in Autism

Total scores from the four instruments that measure neurocognitive development—PPVT, RPCM, SB-5, and VABS—were associated with multiple loci. For SB-5, NVIQ and VIQ scores were associated with four and two independent loci, respectively, while FSIQ was not associated with any locus. The intronic region of the *ACSS3* gene showed the strongest signal for NVIQ. Acyl-CoA synthetase short-chain family member 3 (ACSS3) is a mitochondrial enzyme producing acetyl-CoA from short chain fatty acids., which is necessary for energy creation.(44) NVIQ score was higher for the individuals with TT genotype of the intronic SNV in *ACSS3* (rs7487040, $P = 2.14\times10^{-8}$) (**Figure 4A**). The *ACSS3* gene appears frequently in literature regarding psychiatric disorders, ranging from ADHD to schizophrenia.(45,46) A whole-exome sequencing (WES) study of families with single or multiple ADHD cases uncovered a rare variant in the *ACSS3* gene.(45) Further, a GWAS that incorporated individuals with schizophrenia, bipolar disorder, and schizoaffective disorder revealed an association between a SNP in the *ACSS3* gene (rs7136590, $P = 7.43\times10^{-6}$) and pars orbitalis volume.(47) The pars orbitalis is part of the inferior frontal gyrus and noteworthy to our study of ASD due to its importance for the brain's language processing network.(48,49)

Two loci that were significant for VIQ were mapped to the *PLA2G4A* and *CDH23* genes. Individuals with GG genotype of chr1:186860544 in *PLA2G4A* were associated with higher VIQ scores (**Figure 4B**). The *PLA2G4A* gene encodes the cytosolic phospholipase A2 (PLA2) that plays important for normal brain development and synaptic function.(50) Cytosolic phospholipase A2 (PLA2G4A) is an enzyme that facilitates phospholipid hydrolysis to cleave and, thus, release fatty acids including as arachidonic acid.(50) An earlier GWAS revealed SNPs in *PLA2G4A* associated with epilepsy,(51) a condition that is estimated to be comorbid with ASD in up to 39% of cases.(3) The *CDH23* gene encodes atypical cadherin that is implicated in Usher syndrome type 1, non-syndromic and age-related hearing loss, pre-pulse inhibition, and Alzheimer's disease.(52)

Although the PPVT is a receptive vocabulary test and, therefore, a proxy for verbal intelligence, the SB-5 VIQ and PPVT standard scores were associated with different loci in our cohort. Two intronic loci in the *GDPD4* and *OSTN* genes were significantly associated with the PPVT score (**Figure 3B**). Inherited or *de novo* CNVs encompassing *GDPD4* were found in patients with ASD.(53) Interestingly, the *OSTN* gene, encoding osteocrin, restricts activity-dependent dendritic growth in human neurons. In response to sensory input, osteocrin regulates features of neuronal structure and function that are unique to primates.(54) An additional SNV in the *OSTN* gene was associated with RBS Total score. Additionally, two intergenic loci were significantly associated with the RPCM score, which is a "paradigmatic" measure of fluid intelligence.(55) The total VABS score, which assesses adaptive behavior, was associated with a locus in the *NUGGC* gene.

We discovered that eight loci were significantly associated with HC. These loci were mapped to *CHD5, GRP137B, NKAIN3, UBASH3B,* and intergenic regions. The strongest signal was found for the *NKAIN3* gene (**Figure 4C**), which encodes the Sodium/Potassium Transporting ATPase Interacting 3 protein. *NKAIN3* encompasses a risk allele for dyslexia(56) and is a known candidate gene for Dravet syndrome (MIM# 607208), which is a disorder characterized by an infantile-onset epileptic encephalopathy, intellectual disability, and refractory seizures.(57)

Proportion of phenotype variation explained by genotype

We performed a VCA to estimate the contribution of various covariates– age at test, gender, genetic ancestry, and PRS– in the phenotypic scores (**Additional File 6**). Except for HC for which the covariates accounted for 71.4% of variance in the observed values, an average of 11.5% (range from 0.92% to 44.18%) of phenotypic variance was attributed to the covariates. For HC, age was the major contributor (54.23%) to the phenotype variance, followed by gender (14.2%). In fact, gender and age at test were the most frequent top contributing covariates for several phenotypes– 12 out of 29 (age) and 15 out of 29 (gender)– as well as the largest contributors (5.6% on average by age (ranges from 0 to 38.10%) and 3.4% by gender (ranges from 0 to 14.29%)). Overall, PRS for ASD was not likely attributable to the variance of EPs.

Discussion

Independent studies on multiple cohorts have demonstrated that ASD is highly heritable with genetic underpinnings that are likely polygenic from common and rare variants. Using WGS, previous studies discovered candidate genes with *de novo* mutations and rare inherited variants enriched in individuals with ASD. Gene discovery efforts with genotyping microarrays, WES, and WGS have been successful to catalogue candidate genes for ASD. However, there are still specific genes, molecular mechanisms, and brain circuits implicated in the disorder that are yet to be discovered. More importantly, understanding the biological substrates that underlie specific symptoms will be valuable to define target symptoms for treatments and, thus, to develop therapeutic approaches. To this end, we aimed to discover the genetic basis of the core symptom domains and neurocognitive development in ASD using rich phenotype information and WGS data from the AGRE. All of phenotype tests that we used in the analysis were associated with a genetic locus or multiple loci except for SRS Total Score.

The most significant association was found for ADOS Behavior Total Score (Module 3) and the locus including exon 4 and intronic region of the *PTPRD* gene. For an intronic SNV of *PTPRD*, the scores were higher in individuals of AA genotype of rs12006270, with females displaying more severe behavioral deficits. The *PTPRD* gene encodes the receptor protein tyrosine phosphatase delta (PTPRD) that regulates neurogenesis by modulating tyrosine kinase signaling pathway.(58) Previously, a homozygous microdeletion of this gene was found in a patient with intellectual disability, hearing loss, and trigonocephaly.(59) Decreased dosage of *PTPRD* showed aberrant neurogenesis and an increased number of cortical neurons *in vivo* that suggest *PTPRD* is a key regulator of brain development.(58) Indeed, independent studies have reported genetic association of *PTPRD* with ASD,(60) ADHD,(61)and Obsessive Compulsive Disorder.(62) Moreover, neurofibrillary tangle accumulation in autopsy brain samples from Alzheimer's disease was associated with the *PTPRD* locus (rs560380, $P = 3.8\times10^{-8}$).(63)

It is compelling that 132 loci were significantly associated with RBS Total Score, and 54 of these loci had lead variants with $P < 1.72\times10^{-9}$. RRBs comprise one of the core symptom domains of ASD in the DSM-IV; however, these behaviors are observed in multiple neuropsychiatric conditions (e.g., schizophrenia, bipolar disorder, obsessive-compulsive disorder, drug addiction, L-DOPA-induced dyskinesia, and Huntington's disease).(64) Behavioral approaches are used to treat RRBs and several pharmacological treatments have been effective in reducing these behaviors in ASD. Therefore, RRBs are treatment targets; however, biological pathways and neural circuits associated with RRBs remain undiscovered. Interestingly, the eight loci that were associated with RBS Total Score were enriched with the genes involved in the calcium signaling pathway and highly expressed in various regions of the brain. Parvalbumin (PV) is a calcium binding protein that is expressed in a subpopulation of neurons called fast-spiking interneurons (i.e., PV+ interneurons). PV+ interneurons are reduced in the prefrontal cortex of ASD compared to controls.(65) Moreover, a recent study found that dysregulation of calcium signaling in astrocytes of striatal microcircuits contributed to repetitive behaviors *in vivo*.(66) PV knockout mice exhibit RRBs as well as the other core symptoms of ASD.(67)

Multiple SNVs in coding and intronic regions of the *OSTN* gene were significant for the total scores on the PPVT and RBS. Osteocrin, which is a secretory peptide of 103 amino acids, binds to natriuretic peptide clearance receptor.(68) Osteocrin is involved in activity-dependent regulation of neuronal function, bone growth, and physical endurance.(69) The *OSTN* gene is highly expressed in multiple areas of the developing brain, especially in the neocortex, and shows higher levels of expression in cortical regions compared to the other tissue types in the adult human. Ataman and colleagues identified that osteocrin was secreted in an activity-dependent manner in human fetal brain cultures, but not in mice.(54) Evolutionary acquisition of the regulatory region of *OSTN* enables the binding of MEF2, an activity-regulated transcription factor. As a result, activity-dependent dendritic growth is restricted in human neurons. Indeed, integrative analysis of ChIP-seq, transcriptome, and protein-protein interaction data demonstrated that MEF2A and MEF2C binding sites were enriched in the regulatory regions of ASD candidate genes.(70) In our analysis, an intronic variant (rs6783287, $P=7.4\times10^{-11}$) was significant for a phenotype score related to RRB. Scores on the PPVT and RBS were significantly associated with coding and intronic variants in the *OSTN* gene located at chromosome 3q28. This region is also a GWAS hot spot for cerebrospinal fluid tau levels in Alzheimer's disease,(71) allowing for the conclusion that its association with a detriment in VIQ is indicative of cognitive decline. Since osteocrin is a circulating peptide, it has the potential to be a biomarker of ASD, endophenotype marker for RRBs and verbal intelligence, and potential treatment target for ASD. However, further *in vivo* studies are required to understand downstream biological pathways in human cells.

Limitations

Firstly, all the loci discovered for phenotype scores need to be reproduced in the other cohorts. The AGRE participants are primarily multiplex families with pervasive developmental disorder (PDD) and Asperger syndrome that were diagnosed by experts using the ADI-R and ADOS. Multiplex families with ASD can have higher genetic burden compared to sporadic cases; however, the aim of our analysis was to find genetic substrates of phenotype tests covering core symptom domains and neurocognitive development rather than to discover associations between ASD and neurotypical controls. A similar study can be performed for different cohorts to validate the associations from the current study. Secondly, sample sizes were moderate to discover loci with small effect sizes. For instance, ADOS Module 2 scores were available for a subgroup of our cohort (N=311) while 1,881 individual scores were available for the social and behavior domains of ADI-R. Thirdly, genotype-phenotype associations found in our study may be valid for ASD and their family members. Indeed, the candidate genes with alleles that were previously reported for intelligence were mapped to the genes associated with diverse phenotype scores—ADI-R Nonverbal Communication Total Score (*CADM2*), Behavior Total Score of ADOS Module 3 (*PTPRD* and *GDA*), HC (*LNPEP*) and RBS (*FAM78B, CNTN4, FHIT, ICA1, DGKB, SP4, SGCZ, CDH2, PLCB1, MACROD2*, and *PTPRT*), but not with cognitive measurements such as SB-5 FSIQ and PPVT. As unaffected siblings were included in the analysis, some associations with ADI-R and ADOS scores might indicate the genotype difference between affected and unaffected individuals.

Conclusion

In summary, we used WGS and phenotype scores to successfully perform an endophenotype-wide association analysis that extends previous candidate gene discovery for ASD by unveiling the genetic basis of core symptoms and neurocognitive deficits. Notably, several ASD candidate genes that were previously discovered by case-control comparisons were associated with the severity of core symptoms such as RRBs in the present study. It is possible, therefore, that these candidate genes are responsible for specific traits that constitute core symptoms of ASD rather than the disorder itself. Several genes (such as *OSTN*) identified in our results represent potential biomarkers for ASD; however, further studies are required to replicate our findings and to understand the genetic impacts on molecular pathways, brain circuits, and the phenotype spectrum in the context of RDoC framework.

List of abbreviations

ADHD: Attention deficit hyperactivity disorder

ADI-R: Autism Diagnostic Interview- Revised

ADOS: Autism Diagnostic Observation Schedule

AGRE: Autism Genetic Resource Exchange

ASD: Autism spectrum disorder

CNVs: Copy number variants

DSM: Diagnostic and Statistical Manual of Mental Disorders

EPs: Endophenotypes

FSIQ: Full scale IQ

FUMA: Functional Mapping and Annotation

GWA: Genome-wide association

GWASs: Genome-wide association studies

HC: Head circumference

iHART: Hartwell Autism Research and Technology Initiative

IQ: Intelligence quotient

NQA: Not quite autism

NVIQ: Non-verbal IQ

PCs: Principal components

PPVT: Peabody Picture Vocabulary Test

PRS: Polygenic risk score

PV: Parvalbumin

RBS: Repetitive Behavior Scale

RDoC: Research Domain Criteria

RPCM: Raven's Progressive Colored Matrices

RRBs: Restricted, repetitive behaviors

SB-5: Stanford Binet-5

SD: Standard deviation

SNP: Single nucleotide polymorphism

SNVs: Single nucleotide variants

SRS: Social Responsiveness Scale

VABS: Vineland Adaptive Behavior Scale

VCA: Variance component analysis

VIQ: Verbal IQ

WES: Whole exome sequencing

WGS: Whole genome sequencing

Acknowledgements

We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families.

Ethics approval

The study was approved by the Institutional Review Board of Boston Children's Hospital (IRB-P00020603).

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the AGRE, MSSNG and iHART database.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was support by the grants from the National Institute of Health (R01MH107205, R24OD024622, and U01TR002623).

Author's contributions

IL and EK collected and analyzed the data and wrote the manuscript. SK designed the study, collected, and analyzed the data, and wrote the manuscript. The authors read and approved the manuscript.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. Washington, DC; 2013.

2. Lefort-Besnard J, Vogeley K, Schilbach L, Varoquaux G, Thirion B, Dumas G, et al. Patterns of autism symptoms: hidden structure in the ADOS and ADI-R instruments. Transl Psychiatry. 2020 Jul 30;10(1):257.

3. Jeste SS, Geschwind DH. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nat Rev Neurol. 2014/01/28 ed. 2014 Feb;10(2):74–81.

4. Maenner MJ, Shaw KA, Baio J, EdS1, Washington A, Patrick M, et al. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. Morb Mortal Wkly Rep Surveill Summ Wash DC 2002. 2020 Mar 27;69(4):1–12.

5. Volkmar F, Siegel M, Woodbury-Smith M, King B, McCracken J, State M. Practice Parameter for the Assessment and Treatment of Children and Adolescents With Autism Spectrum Disorder. J Am Acad Child Adolesc Psychiatry. 2014 Feb 1;53(2):237–57.

6. Tick B, Bolton P, Happé F, Rutter M, Rijsdijk F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. J Child Psychol Psychiatry. 2016 May 1;57(5):585–95.

7. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The Familial Risk of Autism. JAMA. 2014 May 7;311(17):1770–7.

8. Bolton P, Macdonald H, Pickles A, Rios P, Goode S, Crowson M, et al. A case-control family history study of autism. J Child Psychol Psychiatry. 1994 Jul;35(5):877–900.

9. Happé F, Ronald A. The ‘Fractionable Autism Triad’: A Review of Evidence from Behavioural, Genetic, Cognitive and Neural Research. Neuropsychol Rev. 2008 Dec 1;18(4):287–304.

10. Lord C, Petkova E, Hus V, Gan W, Lu F, Martin DM, et al. A Multisite Study of the Clinical Diagnosis of Different Autism Spectrum Disorders. Arch Gen Psychiatry. 2012 Mar 1;69(3):306–13.

11. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. Am J Psychiatry. 2010 Jul 1;167(7):748–51.

12. Kendler KS, Neale MC. Endophenotype: a conceptual analysis. Mol Psychiatry. 2010 Aug 1;15(8):789–97.

13. Preston GA, Weinberger DR. Intermediate phenotypes in schizophrenia: a selective review. Dialogues Clin Neurosci. 2005;7(2):165–79.

14. Bodfish JW, Symons FJ, Lewis MH. The repetitive behavior scale. Western Carolina Center Research Reports; 1999.

15. Constantino JN, Gruber CP. Social responsiveness scale: SRS-2. Torrance, CA: Western Psychological Services; 2012.

16. Dunn LM, Dunn LM. Peabody picture vocabulary test-III. Circle Pines, MN: American Guidance Service; 1997.

17. Raven JC, Court JH, Raven J. Manual for Raven’s Progressive Matrices and Vocabulary Scales. Oxford: Oxford Psychologists Press; 1995.

18. Roid GH, Pomplun M. The Stanford-Binet Intelligence Scales, Fifth Edition. In: Contemporary intellectual assessment: Theories, tests, and issues, 3rd ed. New York, NY, US: The Guilford Press; 2012. p. 249–68.

19. Sparrow SS, Balla DA, Cicchetti DV, Harrison PL. Vineland adaptive behavior scales. Circle Pines, MN: American Guidance Service; 1984.

20. Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, et al. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. Am J Hum Genet. 2001 Aug;69(2):463–6.

21. Lord C, Rutter M, Le Couteur 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. 1994 Oct 1;24(5):659–85.

22. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, et al. The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. J Autism Dev Disord. 2000 Jun 1;30(3):205–23.

23. C Yuen RK, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017 Apr 1;20(4):602–11.

24. Ruzzo EK, Pérez-Cano L, Jung J-Y, Wang L, Kashef-Haghighi D, Hartl C, et al. Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. Cell. 2019 Aug 8;178(4):850-866.e26.

25. Sacco R, Curatolo P, Manzi B, Militerni R, Bravaccio C, Frolli A, et al. Principal pathogenetic components and biological endophenotypes in autism spectrum disorders. Autism Res. 2010 Oct 1;3(5):237–52.

26. Mottron L. Matching Strategies in Cognitive Research with Individuals with High-Functioning Autism: Current Practices, Instrument Biases, and Recommendations. J Autism Dev Disord. 2004 Feb;34(1):19–27.

27. Engel de Abreu PMJ, Conway ARA, Gathercole SE. Working memory and fluid intelligence in young children. Intelligence. 2010 Nov 1;38(6):552–61.

28. Sparrow SS, Balla DA, Cicchetti DV, Harrison PL. Vineland adaptive behavior scales. Circle Pines, MN: American Guidance Service; 1984.

29. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol. 2008 May 1;32(4):381–5.

30. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017 Nov 28;8(1):1826.

31. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019 Mar 1;51(3):431–44.

32. Schildkraut JJ. The Catecholamine Hypothesis of Affective Disorders: A Review of Supporting Evidence. Am J Psychiatry. 1965 Nov 1;122(5):509–22.

33. Cases-Langhoff C, Voss B, Garner A, Appeltauer U, Takei K, Kindler S, et al. Piccolo, a novel 420 kDa protein associated with the presynaptic cytomatrix. Eur J Cell Biol. 1996 Mar;69(3):214—223.

34. Need AC, McEvoy JP, Gennarelli M, Heinzen EL, Ge D, Maia JM, et al. Exome Sequencing Followed by Large-Scale Genotyping Suggests a Limited Role for Moderately Rare Risk Factors of Strong Effect in Schizophrenia. Am J Hum Genet. 2012 Aug 10;91(2):303–12.

35. van der Zwaag B, Franke L, Poot M, Hochstenbach R, Spierenburg HA, Vorstman JAS, et al. Gene-Network Analysis Identifies Susceptibility Genes Related to Glycobiology in Autism. PLOS ONE. 2009 May 28;4(5):e5324.

36. Hartshorne TS, Grialou TL, Parker KR. Autistic-like behavior in CHARGE syndrome. Am J Med Genet A. 2005 Mar 15;133A(3):257–61.

37. Chauvet S, Cohen S, Yoshida Y, Fekrane L, Livet J, Gayet O, et al. Gating of Sema3E/PlexinD1 Signaling by Neuropilin-1 Switches Axonal Repulsion to Attraction during Brain Development. Neuron. 2007 Dec 6;56(5):807–22.

38. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013 Oct 3;4(1):36–36.

39. Hidenori Ito, Morishita R, Mizuno M, Kawamura N, Tabata H, Nagata K-I. Biochemical and Morphological Characterization of a Neurodevelopmental Disorder-Related Mono-ADP-Ribosylhydrolase, MACRO Domain Containing 2. Dev Neurosci. 2018 Sep 18;40(3):278–87.

40. Autism Spectrum Disorders Working Group of the Psychiatric Genomics Consortium. Meta-analysis of GWAS of overr 16,000 individuals wih autism specttrum disorder highlights a novel locus at 10q24.32 and a signfiicant overlap wih schizophrenia. Mol Autism. 2017 May 22;8(21).

41. Rachel Jones, Cadby G, Blangero J, Abraham L, Whitehouse A, Moses E. MACROD2 gene associated witth autisttic-like trarits in a general populattion sample. Psychiatry Genet. 2014 Dec;24(6):241–8.

42. Lionel A, Crosbie J, Barbosa N, Goodale T, Thiruvahindrapuram B, Rickaby J, et al. Rare copy number variation discovery and cross-disorder comparisions idenitfy risk genes for ADHD. Sci Transl Med. 2011 Aug 10;3(95).

43. Maas N, Van de Putte T, Melotte C, Francis A, Constance Schrander-Stumpel, Sanlaville D, et al. The C20orf133 gene is disrupted in a patientt with Kabuki syndrome. J Med Genet. 2007 Sep;44(9):562–9.

44. Hunter AM, Leuchter AF, Power RA, Muthén B, McGrath PJ, Lewis CM, et al. A genome-wide association study of a sustained pattern of antidepressant response. J Psychiatr Res. 2013 Sep 1;47(9):1157–65.

45. Al-Mubarak BR, Omar A, Baz B, Al-Abdulaziz B, Magrashi AI, Al-Yemni E, et al. Whole exome sequencing in ADHD trios from single and multi-incident families implicates new candidate genes and highlights polygenic transmission. Eur J Hum Genet. 2020 Aug 1;28(8):1098–110.

46. Manchia M, Piras IS, Huentelman MJ, Pinna F, Zai CC, Kennedy JL, et al. Pattern of gene expression in different stages of schizophrenia: Down-regulation of NPTX2 gene revealed by a meta-analysis of microarray datasets. Eur Neuropsychopharmacol. 2017 Oct 1;27(10):1054–63.

47. Alliey-Rodriguez N, Grey TA, Shafee R, Asif H, Lutz O, Bolo NR, et al. NRXN1 is associated with enlargement of the temporal horns of the lateral ventricles in psychosis. Transl Psychiatry. 2019 Sep 17;9(1):230–230.

48. Petrides M. Cytoarchitecture. In: Neuroanatomy of Language Regions of the Human Brain [Internet]. San Diego: Academic Press; 2014. p. 89–138. Available from: https://www.sciencedirect.com/science/article/pii/B9780124055148500050

49. Webb WG. 9 - Central Language Mechanism and Learning. In: Webb WG, editor. Neurology for the Speech-Language Pathologist (Sixth Edition) [Internet]. Mosby; 2017. p. 181–205. Available from: https://www.sciencedirect.com/science/article/pii/B9780323100274000099

50. Negre-Aminou P, Pfenninger KH. Arachidonic Acid Turnover and Phospholipase A2 Activity in Neuronal Growth Cones. J Neurochem. 1993 Mar 1;60(3):1126–36.

51. EPICURE Consortium, EMINet Consortium, Steffens M, Leu C, Ruppert A-K, Zara F, et al. Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. Hum Mol Genet. 2012 Dec 15;21(24):5359–72.

52. Bork JM, Peters LM, Riazuddin S, Bernstein SL, Ahmed ZM, Ness SL, et al. Usher Syndrome 1D and Nonsyndromic Autosomal Recessive Deafness DFNB12 Are Caused by Allelic Mutations of the Novel Cadherin-Like Gene CDH23. Am J Hum Genet. 2001 Jan 1;68(1):26–37.

53. Bucan M. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. PLoS Genet. 2009 Jun 26;5(6).

54. Ataman B, Boulting GL, Harmin DA, Yang MG, Baker-Salisbury M, Yap E-L, et al. Evolution of Osteocrin as an activity-regulated factor in the primate brain. Nature. 2016 Nov 10;539(7628):242–7.

55. Sutherland S. Cognition: Parallel distributed processing. Nature. 1986 Oct 1;323(6088):486–486.

56. Gialluisi A, Andlauer TFM, Mirza-Schreiber N, Moll K, Becker J, Hoffmann P, et al. Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. Transl Psychiatry. 2019 Feb 11;9(1):77.

57. Carvill GL, Weckhuysen S, McMahon JM, Hartmann C, Møller RS, Hjalgrim H, et al. <em>GABRA1</em> and <em>STXBP1</em>: Novel genetic causes of Dravet syndrome. Neurology. 2014 Apr 8;82(14):1245.

58. Tomita H, Cornejo F, Aranda-Pino B, Woodard CL, Rioseco CC, Neel BG, et al. The Protein Tyrosine Phosphatase Receptor Delta Regulates Developmental Neurogenesis. Cell Rep. 2020 Jan 7;30(1):215-228.e5.

59. Choucair N, Mignon-Ravix C, Cacciagli P, Abou Ghoch J, Fawaz A, Mégarbané A, et al. Evidence that homozygous PTPRD gene microdeletion causes trigonocephaly, hearing loss, and intellectual disability. Mol Cytogenet. 2015 Jun 16;8(1):39.

60. Levy D, Ronemus M, Yamrom B, Lee Y, Leotta A, Kendall J, et al. Rare De Novo and Transmitted Copy-Number Variation in Autistic Spectrum Disorders. Neuron. 2011 Jun 9;70(5):886–97.

61. Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, et al. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry. 2010 Jun 1;15(6):637–46.

62. Pauls DL, Abramovitch A, Rauch SL, Geller DA. Obsessive–compulsive disorder: an integrative genetic and neurobiological perspective. Nat Rev Neurosci. 2014 Jun 1;15(6):410–24.

63. Chibnik LB, White CC, Mukherjee S, Raj T, Yu L, Larson EB, et al. Susceptibility to neurofibrillary tangles: role of the PTPRD locus and limited pleiotropy with other neuropathologies. Mol Psychiatry. 2018 Jun 1;23(6):1521–9.

64. Crittenden JR, Gipson TA, Smith AC, Bowden HA, Yildirim F, Fischer KB, et al. Striatal transcriptome changes linked to drug-induced repetitive behaviors. Eur J Neurosci. 2021 Apr 1;53(8):2450–68.

65. Ariza J, Rogers H, Hashemi E, Noctor SC, Martínez-Cerdeño V. The Number of Chandelier and Basket Cells Are Differentially Decreased in Prefrontal Cortex in Autism. Cereb Cortex. 2018 Feb 1;28(2):411–20.

66. Yu X, Taylor AMW, Nagai J, Golshani P, Evans CJ, Coppola G, et al. Reducing Astrocyte Calcium Signaling In Vivo Alters Striatal Microcircuits and Causes Repetitive Behavior. Neuron. 2018 Sep 19;99(6):1170-1187.e9.

67. Wöhr M, Orduz D, Gregory P, Moreno H, Khan U, Vörckel KJ, et al. Lack of parvalbumin in mice leads to behavioral deficits relevant to all human autism core symptoms and related neural morphofunctional abnormalities. Transl Psychiatry. 2015 Mar 1;5(3):e525–e525.

68. Moffatt P, Thomas G, Sellin K, Bessette M-C, Lafrenière F, Akhouayri O, et al. Osteocrin is a specific ligand of the natriuretic Peptide clearance receptor that modulates bone growth. J Biol Chem. 2007 Dec;282(50):36454—36462.

69. Subbotina E, Sierra A, Zhu Z, Gao Z, Koganti SRK, Reyes S, et al. Musclin is an activity-stimulated myokine that enhances physical endurance. Proc Natl Acad Sci U S A. 2015 Dec;112(52):16042—16047.

70. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism. Cell. 2013 Nov 21;155(5):1008–21.

71. Cruchaga C, Kauwe JSK, Harari O, Jin SC, Cai Y, Karch CM, et al. GWAS of Cerebrospinal Fluid Tau Levels Identifies Risk Variants for Alzheimer's Disease. Neuron. 2013 Apr 24;78(2):256–68.

Additional Files

Additional File 1. Summary of demographic information for AGRE dataset.

Additional File 2. Descriptions of phenotype instruments.

Additional File 3. Phenotype measures used for association study and data availability.

Additional File 4. List of genomic loci significantly associated with phenotype scores.

Additional File 5. Regional plots for risk loci associated with phenotype scores (excludes loci shown in main figures).

Additional File 6. Proportion of phenotype variance attributed by covariates – age at test, race, sex, and polygenic risk score.

Figure legends

Figure 1. Overview of genomic loci associated with phenotypic scores. Genome-wide association analysis with each of phenotype scores highlights significant loci and candidate genes for endophenotype. Horizontal axis indicates genomic position from chromosome 1 to chromosome 22 and each row in vertical axis is organized by test instruments and phenotypic scores. The phenotype scores with significantly associated loci are (top to bottom): ADOS Module 1 (social, social/communication and total scores), ADOS Module 3 (communication and behavioral total scores), four calculated total scores on ADI-R (social, nonverbal communication, verbal communication, and behavior), RBS overall score, RPCM total score, PPVT score, SB-5 (VIQ and NVIQ), VABS standard score, and HC. Circles indicate genomic loci passed genome-wide significance ($P < 5 \times 10^{-8}$), where the bigger the size the smaller the nominal P-value. Genomic positions across different chromosomes are indicated by alternating colors between chromosomes (blue – grey), but loci that satisfy the adjusted P-value threshold for multiple (N=29) concurrent hypothesis testing ($P < 1.72 \times 10^{-9}$) are highlighted in red. The genes that overlap with or in flanking region of each significant genomic loci are displayed next to the corresponding circles.

Figure 2. Regional plots for the significant loci associated with domain scores of ADOS and ADI-R. The purple diamond shape indicates the most significant single nucleotide variant (SNV) within a region. The distribution of phenotype scores by genotype of the most significant variant (lead SNV) is shown in an insert. The blue line plot shows the recombination rates (cMM/Mb) across genomic positions. **a** The genomic locus near the 3′ untranslated region of *CDYL* is associated with ADI-R verbal communication score. The individuals with CC genotype for the lead SNV show lower score. The females with CC genotype are unaffected siblings except for one with score > 15. **b** Two genic regions of *PCLO* and *SEMA3E* were found significantly associated to ADI-R nonverbal communication total score. For the lead variant in the *PCLO* gene, individuals with AA genotype showed lower phenotypic score in both males (all unaffected siblings) and females (all probands). Also, the ADI-R non-verbal communication total

score decreased among individuals with GG genotypes for the lead SNV in the *SEMA3E* gene. **c** The *HOXC11* and *HOXC10* loci are significantly associated with ADOS Module 1 total score.

Figure 3. Regional plots for the significant loci associated with RBS and PPVT scores. a Two regions in the *MACROD2* gene are associated with RBS total scores. **b** Two genic regions in the *OSTN* gene are associated with RBS and PPVT scores. The lead variants are different as indicated by the positions of purple diamond shapes. The plots on the right-side shows the change of each score distribution by genotypes of the lead variants. For the lead variant of RBS score, all individuals with GG genotype were probands (1 female and 2 males).

Figure 4. Regional plots for the significant loci associated with SB-5 nonverbal and verbal IQ scores and head circumference. a and **b** show the genomic loci in the *ACSS3* and *PLA2G4A* genes associated with SB-5 nonverbal and verbal IQ scores, respectively. **c** Head circumference is significantly associated with the locus close to 5´- untranslated region of *NKAIN3*.

Table 1. Genomic loci associated with severity of ASD and neurocognitive measurements.

Phenotype Measures	Name	Locus on GRCh37	rsID for lead variant	gene mapped to lead variant	BETA	SE	*P*-value for lead variant
ADI-R	social total	chr10:3535233-3622867	rs630310	*PFKP; KLF6*	-1.88	0.341	4.39E-08
	verbal communication total	chr6:4940289-5054132	rs11754469	*CDYL* (intron)	-3.42	0.550	7.15E-10
	behavior total	chr7:43450417-43480484	rs7794675	*HECW1* (intron)	1.01	0.174	7.97E-09
	non-verbal communication total	chr7:82409046-82694115	rs73710023	*PCLO* (intron)	-3.27	0.528	1.28E-09
	non-verbal communication total	chr7:83020916-83130343	rs3109789	*SEMA3E* (intron)	-0.83	0.145	1.99E-08
ADOS	Module 1 total	chr12:54355209-54380016	rs35493008	*HOXC11; HOXC10*	-1.91	0.336	2.16E-08
	Module 3 - behavior total	chr8:95090293-95124797	rs9987124	*CDH17* (downstream)	1.59	0.276	1.45E-08
	Module 3 - behavior total	chr9:10029787-10042270	rs12006270	*PTPRD* (intron)	0.69	0.124	3.38E-08
Repetitive Behavior Scale-Revised	Overall score	chr2:15308273-15460222	rs11681691	*NBAS* (intron)	28.03	4.389	3.47E-10
		chr2:153248346-153405829	rs2678296	*FMNL2* (intron)	23.07	3.731	1.18E-09
		chr3:38660936-38669228	rs12498069	*SCN5A* (intron)	20.65	3.293	7.00E-10
		chr3:190922618-190922618	rs6783287	*OSTN* (intron)	39.78	5.994	7.40E-10
		chr4:139266308-139305143	rs6841249	*LINC00499* (intron)	14.50	2.118	1.89E-11
		chr5:94345227-94388511	rs12514324	*MCTP1* (intron)	34.56	5.206	7.25E-11
		chr7:18638652-18750718	rs151012554	*HDAC9* (intron)	40.01	5.987	5.43E-11
		chr7:24548327-24831435	rs17149958	*DFNA5* (intron)	28.28	3.989	3.87E-12
		chr7:29956425-29996459	rs17158543	*SCRN1* (intron)	38.30	5.158	3.99E-13
		chr10:113888983-113950637	rs755792809	*GPAM* (intron)	19.76	3.526	3.22E-08
		chr11:13908345-13976643	rs76115555	*RP11-231N3.1* (intron)	26.09	4.076	3.15E-10
		chr13:38964218-39111269	rs9532234	*UFM1* (downstream)	14.58	2.442	4.11E-09
		chr14:38683261-38708907	rs10498340	*SSTR1; CLEC14A*	23.59	3.809	1.12E-09
		chr14:39502539-39863907	rs2274398	*CTAGE5* (intron)	29.60	5.039	7.07E-09
		chr16:72999435-73023079	rs11640825	*ZFHX3* (intron)	9.19	1.469	7.78E-10
		chr16:81224495-81264177	rs4077825	*PKD1L2* (intron)	12.36	2.094	6.11E-09
		chr18:26102060-26154719	rs73946295	*CDH2* (upstream)	33.36	5.391	1.14E-09
		chr20:15721596-15737935	rs6043502	*MACROD2* (intron)	18.19	2.658	1.93E-11
		chr20:3840539-4028706	rs11906179	*PANK2* (upstream)	23.60	3.514	4.46E-11
Peabody Picture Vocabulary Test	Standard score	chr3:190953758-190975700	rs12498038	*OSTN* (intron)	-9.94	1.728	1.35E-08
		chr11:76924440-77196354	rs11237142	*GDPD4* (intron)	-11.53	2.071	3.79E-08
Stanford-Binet IQ	Verbal IQ	chr1:186859675-186934187	rs10708367	*PLA2G4A* (intron)	9.24	1.557	6.28E-09
		chr10:73454907-73457917	rs1227073	*CDH23* (intron)	9.29	1.662	4.19E-08
	Non-verbal IQ	chr12:81543536-81679597	rs7487040	*ACSS3* (intron)	-9.14	1.600	2.14E-08
Head circumference		chr1:6145023-6180986	rs749435	*KCNAB2* (intron)	1.26	0.227	3.43E-08
		chr8:63158039-63581067	rs9792368	*NKAIN3* (intron)	0.56	0.085	3.38E-11

Figures

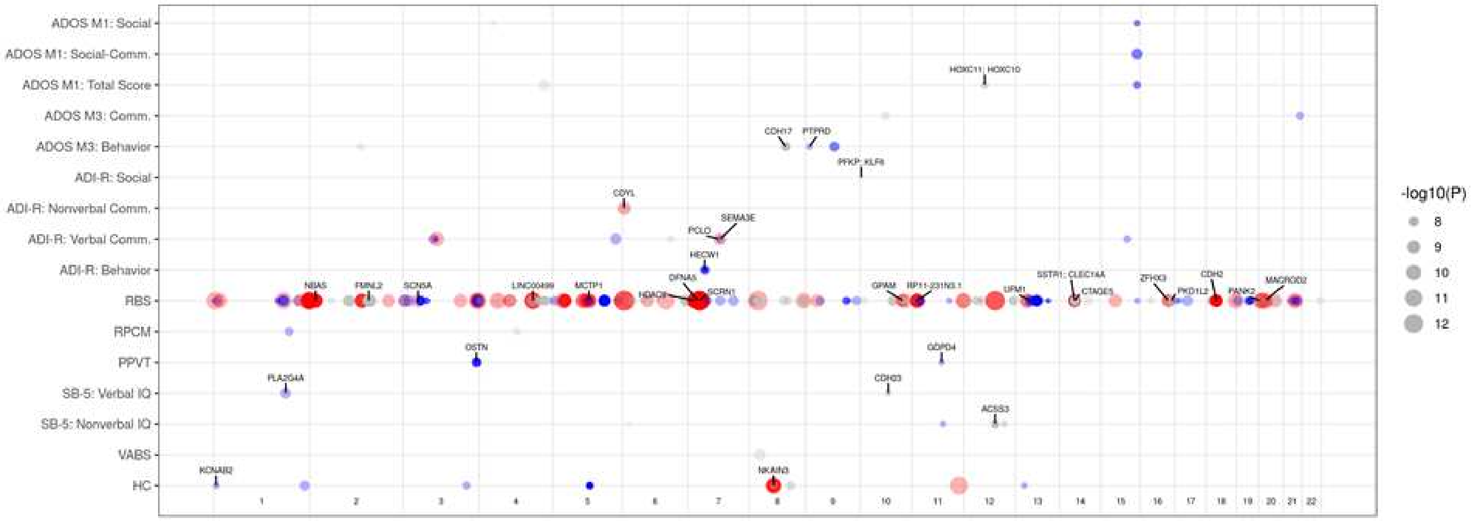

Figure 1

Overview of genomic loci associated with phenotypic scores. Genome-wide association analysis with each of phenotype scores highlights significant loci and candidate genes for endophenotype. Horizontal axis indicates genomic position from chromosome 1 to chromosome 22 and each row in vertical axis is organized by test instruments and phenotypic scores. The phenotype scores with significantly associated loci are (top to bottom): ADOS Module 1 (social, social/communication and total scores), ADOS Module 3 (communication and behavioral total scores), four calculated total scores on ADI-R (social, nonverbal communication, verbal communication, and behavior), RBS overall score, RPCM total score, PPVT score, SB-5 (VIQ and NVIQ), VABS standard score, and HC. Circles indicate genomic loci passed genome-wide significance (P <5 × 10-8), where the bigger the size the smaller the nominal P-value. Genomic positions across different chromosomes are indicated by alternating colors between chromosomes (blue – grey), but loci that satisfy the adjusted P-value threshold for multiple (N=29) concurrent hypothesis testing (P

<1.72 × 10-9) are highlighted in red. The genes that overlap with or in flanking region of each significant genomic loci are displayed next to the corresponding circles.

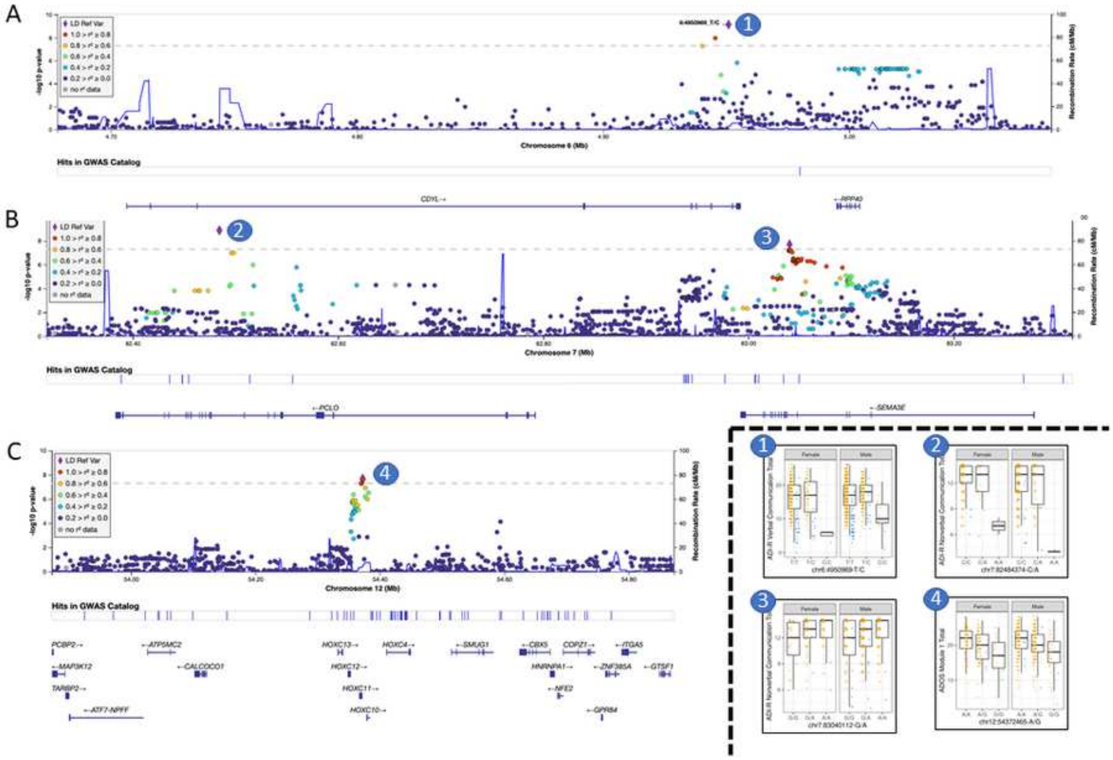

Figure 2

Regional plots for the significant loci associated with domain scores of ADOS and ADI-R. The purple diamond shape indicates the most significant single nucleotide variant (SNV) within a region. The distribution of phenotype scores by genotype of the most significant variant (lead SNV) is shown in an insert. The blue line plot shows the recombination rates (cMM/Mb) across genomic positions. a The genomic locus near the 3´ untranslated region of CDYL is associated with ADI-R verbal communication score. The individuals with CC genotype for the lead SNV show lower score. The females with CC genotype are unaffected siblings except for one with score > 15. b Two genic regions of PCLO and SEMA3E were found significantly associated to ADI-R nonverbal communication total score. For the lead variant in the PCLO gene, individuals with AA genotype showed lower phenotypic score in both males (all unaffected siblings) and females (all probands). Also, the ADI-R non-verbal communication total score decreased among individuals with GG genotypes for the lead SNV in the SEMA3E gene. c The HOXC11 and HOXC10 loci are significantly associated with ADOS Module 1 total score.

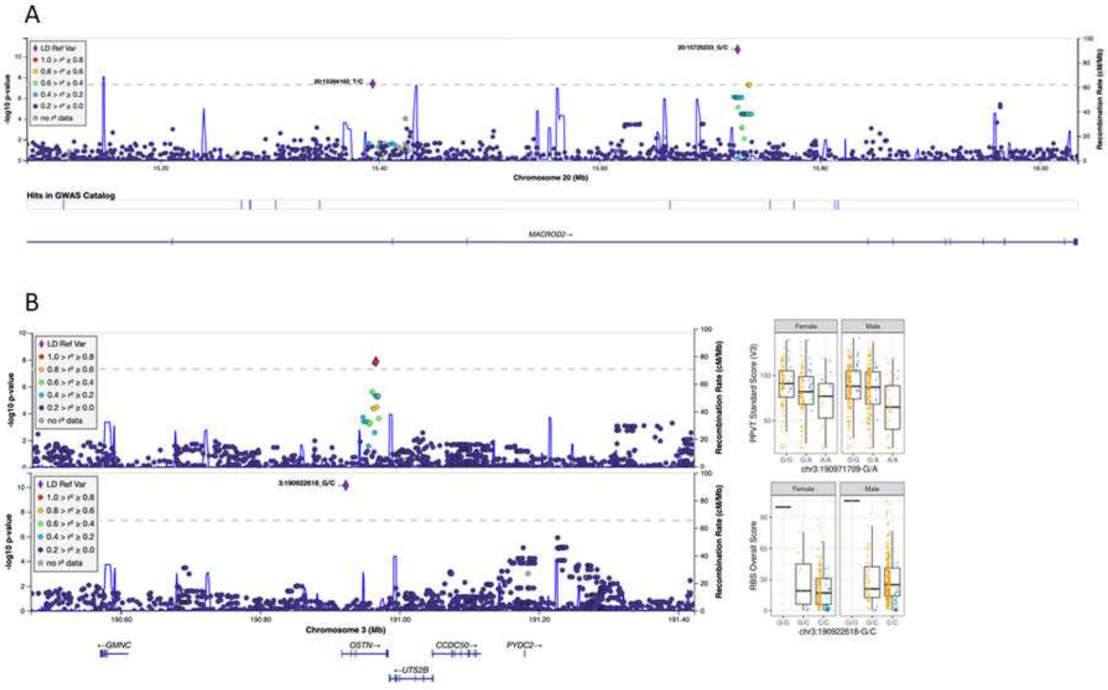

Figure 3

Regional plots for the significant loci associated with RBS and PPVT scores. a Two regions in the MACROD2 gene are associated with RBS total scores. b Two genic regions in the OSTN gene are associated with RBS and PPVT scores. The lead variants are different as indicated by the positions of purple diamond shapes. The plots on the right-side shows the change of each score distribution by genotypes of the lead variants. For the lead variant of RBS score, all individuals with GG genotype were probands (1 female and 2 males).

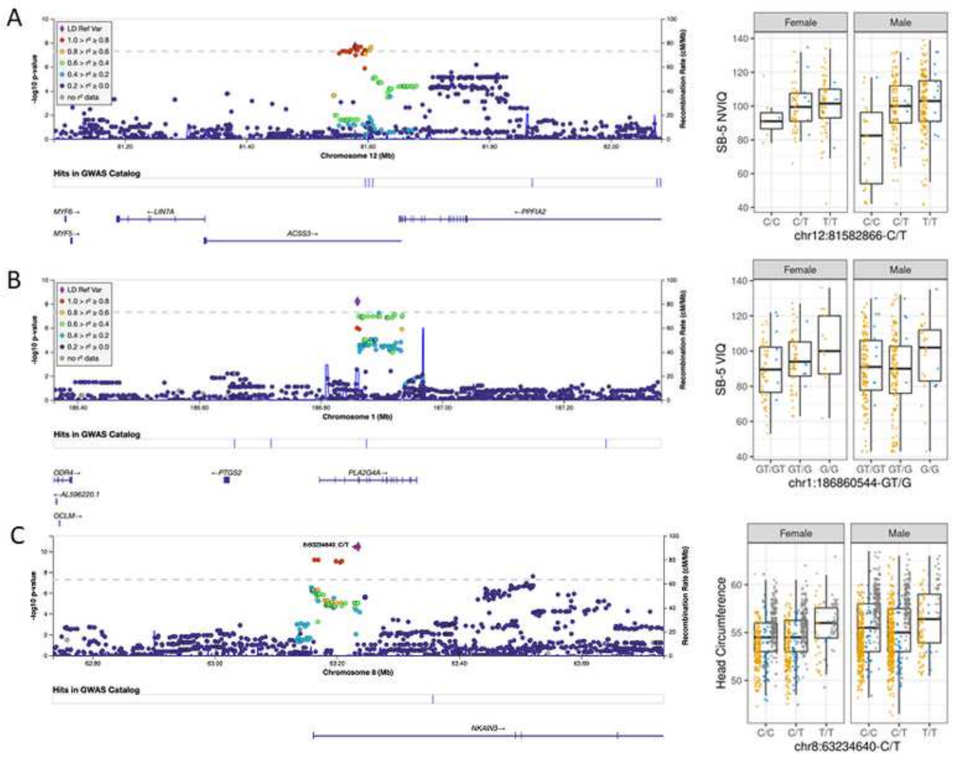

Figure 4

Regional plots for the significant loci associated with SB-5 nonverbal and verbal IQ scores and head circumference. a and b show the genomic loci in the ACSS3 and PLA2G4A genes associated with SB-5 nonverbal and verbal IQ scores, respectively. c Head circumference is significantly associated with the locus close to 5´- untranslated region of NKAIN3.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.pdf
- AdditionalFile2.pdf
- AdditionalFile3.pdf
- AdditionalFile4.pdf
- AdditionalFile5.pdf
- AdditionalFile6.pdf