Title
Level of Attention to Motherese Speech as an Early Marker of Autism Spectrum Disorder.

Permalink
https://escholarship.org/uc/item/75135220

Journal
JAMA network open, 6(2)

ISSN
2574-3805

Authors
Pierce, Karen
Wen, Teresa H
Zahiri, Javad
et al.

Publication Date
2023-02-01

DOI
10.1001/jamanetworkopen.2022.55125

Peer reviewed

ARTICLES
https://doi.org/10.1038/s41588-022-01064-5

nature genetics

A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex

Danny Antaki[1,2,3,4,5,6,18], James Guevara[2,3,18], Adam X. Maihofer[3], Marieke Klein[2,3], Madhusudan Gujral[2,3], Jakob Grove[7,8,9,10], Caitlin E. Carey[11], Oanh Hong[2,3], Maria J. Arranz[12], Amaia Hervas[13], Christina Corsello[14], Keith K. Vaux[15], Alysson R. Muotri[4,5,16], Lilia M. Iakoucheva[3,5], Eric Courchesne[6,17], Karen Pierce[6,17], Joseph G. Gleeson[5,6], Elise B. Robinson[11], Caroline M. Nievergelt[3] and Jonathan Sebat[2,3,4,5] ✉

The genetic etiology of autism spectrum disorder (ASD) is multifactorial, but how combinations of genetic factors determine risk is unclear. In a large family sample, we show that genetic loads of rare and polygenic risk are inversely correlated in cases and greater in females than in males, consistent with a liability threshold that differs by sex. De novo mutations (DNMs), rare inherited variants and polygenic scores were associated with various dimensions of symptom severity in children and parents. Parental age effects on risk for ASD in offspring were attributable to a combination of genetic mechanisms, including DNMs that accumulate in the paternal germline and inherited risk that influences behavior in parents. Genes implicated by rare variants were enriched in excitatory and inhibitory neurons compared with genes implicated by common variants. Our results suggest that a phenotypic spectrum of ASD is attributable to a spectrum of genetic factors that impact different neurodevelopmental processes.

The major risk factors for ASD are genetic and include a variety of rare and common alleles, including rare de novo copy number variants (CNVs)[1] or protein-truncating SNPs and indels of large effect[2], and common polygenic risk that is measured as the sum of thousands of common alleles with small effects[3]. Despite the success in identifying and characterizing multiple types of genetic risk, there is no one variant, gene or polygenic score (PS) that has a high predictive value for an ASD diagnosis. Even CNVs with large effect sizes (odds ratio (OR) > 30) for ASD present with variable psychiatric traits[4] and risk is attributable to a combination of rare and common variations[5,6].

Sex is also a major genetic factor that influences ASD risk. Males are diagnosed with ASD more frequently than females at a ratio of 4:1. A small proportion of cases are associated with X-linked variants[7], but the male preponderance of ASD is not largely explained by genetic variation on sex chromosomes. We and others have hypothesized that it may instead be explained by sex differences in the effects of autosomal variants[8–10]. This hypothesis is supported by previous studies showing that females with ASD have a greater burden of rare CNVs[1,11,12] and gene mutations[13,14]. However, gene-by-sex interactions in ASD have not been examined systematically.

Previous genetic studies have been focused on defining new categories of rare variant risk from DNA-sequencing or by improving the statistical power of genome-wide association studies (GWASs). How combinations of multiple genetic factors contribute to risk and clinical presentation is not known. In the present study, we investigate, in a large dataset of whole genomes and exomes, the combined contributions of de novo, rare inherited and polygenic risk to ASD. We show that the genetic architecture of ASD varies as a spectrum of rare and common variation, each having distinct phenotypic correlates and differential effects in males and females.

Results

Defining multiple components of genetic risk. We investigated the combined effects of multiple genetic factors, detectable by genome sequencing or a combination of exome sequencing and SNP genotyping, on risk of ASD. We focused on several factors that have established associations with case status, such as de novo protein-truncating (dnLoF) and missense (dnMIS) mutations[1,2] and

[1]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA, USA. [2]Beyster Center for Psychiatric Genomics, University of California San Diego, La Jolla, CA, USA. [3]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA. [4]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA. [5]Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA. [6]Department of Neurosciences, University of California San Diego, La Jolla, CA, USA. [7]Department of Biomedicine, Aarhus University, Aarhus, Denmark. [8]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark. [9]Center for Genomics and Personalized Medicine and Center for Integrative Sequencing, iSEQ, Aarhus, Denmark. [10]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [11]Harvard T.H. Chan School of Public Health, Broad Institute of the Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [12]Research Laboratory Unit, Fundacio Docencia i Recerca Mutua, Terrassa, Spain. [13]Child and Adolescent Mental Health Unit, Hospital Universitari Mútua de Terrassa, Barcelona, Spain. [14]TEACCH Autism Program, University of North Carolina, Chapel Hill, NC, USA. [15]Human Longevity Inc., San Diego, CA, USA. [16]Department of Pediatrics and Department of Cellular & Molecular Medicine, University of California San Diego, School of Medicine, Center for Academic Research and Training in Anthropogeny, Archealization Center, Kavli Institute for Brain and Mind, La Jolla, CA, USA. [17]Autism Center of Excellence, University of California San Diego, La Jolla, CA, USA. [18]These authors contributed equally: Danny Antaki, James Guevara. ✉e-mail: jsebat@ucsd.edu

rare inherited variants[15,16] that disrupt genes (inhLoF) and polygenic scoring models that have been associated with ASD case status, including PSs for ASD (PS_{ASD}), schizophrenia (PS_{SZ}) and educational attainment (PS_{EA})[17,18] (see Methods for details on the selection of genetic factors).

We confirmed genetic associations by whole-genome analysis of 37,375 individuals from 11,313 ASD families (12,270 cases, 5,190 typically developing siblings and 19,917 parents). The sample was composed of three datasets, including whole-genome sequencing (WGS) of cohorts from University of California San Diego (UCSD) (https://sebatlab.org/reach-project) and the Simons Simplex Collection (SSC), and exomes and SNP genotyping from the SPARK study[19] (see Methods and Supplementary Tables 1 and 2). SNPs, indels, structural variants (SVs), DNM calling and calculation of ancestry principal components (PCs) were performed using functionally equivalent pipelines for each dataset as described in Methods, and PSs were calculated using the polygenic scoring method SbayesR[20]. Rare variants were annotated for gene functional constraint. Analysis of protein-coding loss of function (LoF) and *cis*-regulatory (CRE) variants was restricted to variant-intolerant genes (LoF observed/expected upper bound fraction (LOEUF) < 0.37) and analysis of missense variants was restricted to those with missense badness (missense badness, PolyPhen-2 and constraint (MPC)) scores > 2.

Association tests were performed for case–control differences in DNM burden. Association of inherited risk was tested using the transmission disequilibrium test (TDT)[15]. Common variant associations were tested using a polygenic TDT (pTDT) that measures overtransmission of risk alleles as the deviation of the offspring PS from the average PS of the parents[17]. We confirmed that de novo synonymous variants were not associated with ASD in the combined sample (Extended Data Fig. 1a) and rates of DNMs were not influenced by batch effects or other confounders (Extended Data Fig. 1b,c). Results confirm significant contributions from genetic factors, including de novo LoF (dnLoF) and missense (dnMIS) mutations (Fig. 1a and Supplementary Tables 3–6). TDT confirmed the associations of rare inherited protein-truncating SNVs (inhLoF) and SVs (LoFSV) (Fig. 1b and Supplementary Tables 7–9). SVs that disrupt CRE variants (CRE-SVs) of constrained genes showed differential transmission in cases and controls, but the TDT did not reach statistical significance in cases. The PSs for autism (PS_{ASD},), schizophrenia (PS_{SZ}) and educational attainment (PS_{EA}) were all significantly associated with ASD (Fig. 1c and Supplementary Table 10) and the polygenic contribution to ASD was consistent across all three cohorts (Supplementary Table 11).

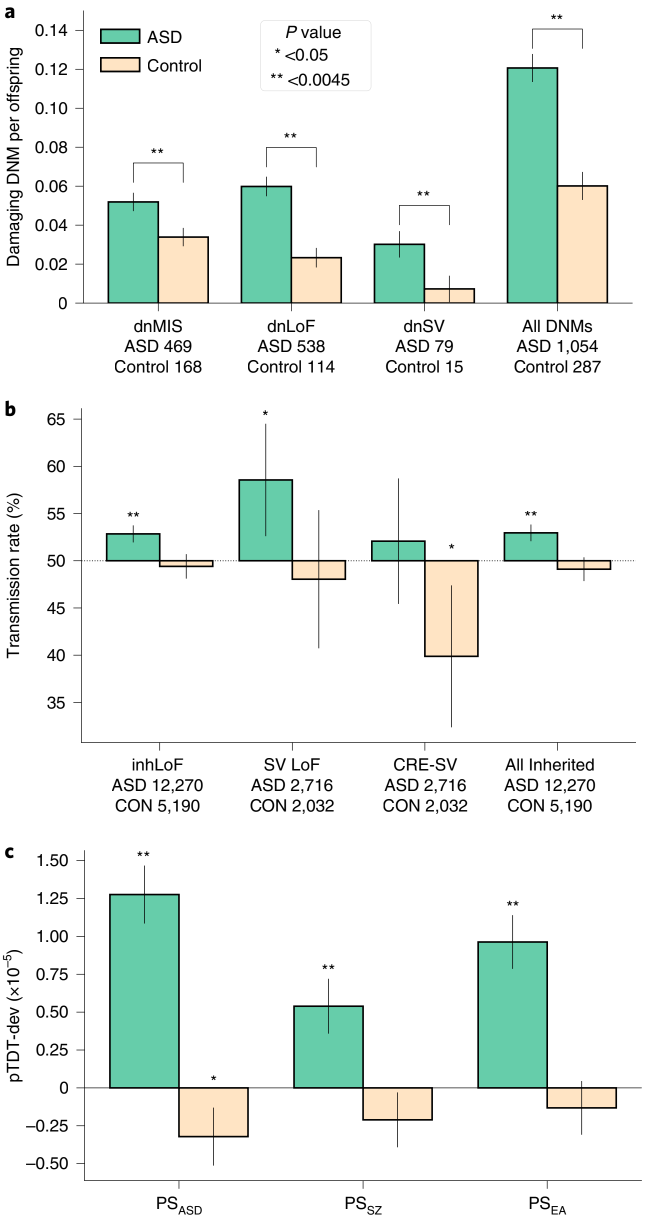


Fig. 1 | Risk for ASD is attributable to multiple genetic factors including DNMs, rare inherited variants and polygenic risk. Multiple genetic factors that have been previously associated with ASD were confirmed in our combined sample. '**' denotes associations that were significant after correction for 11 tests ($P < 0.0045$). Error bars represent the 95% CIs. **a**, Damaging DNMs in genes that are functionally constrained (LOEUF < 0.37 and MPC ≥ 2), including dnMISs and protein-truncating SNVs, indels (dnLoF) and SVs (dnSV), occurring at higher frequencies in cases than in sibling controls. *P* values were based on two-sided Student's *t*-tests. **b**, Protein-truncating SNVs and indels (inhLOF) and SVs (SVLoF), and noncoding SVs that disrupt CRE-SVs, were associated with ASD based on a TDT. **c**, The pTDT was significant for all three PSs, PS_{ASD}, PS_{SZ} and PS_{EA}. Rare variant associations (**a** and **b**) were tested in the full sample ($n = 37{,}375$). The pTDT association was tested in samples of European ancestry ($n = 25{,}391$). Results for **a**–**c** and full lists of rare de novo and inherited variants in constrained genes are provided in Supplementary Tables 3–10.

We examined the combined effects of rare and common variation. To ensure that genetic factors were ascertained consistently across the three cohorts, analysis was restricted to six categories that are detectable in exome and WGS with comparable sensitivity: dnLoF, dnMIS, inhLoF and PSs (PS_{ASD}, PS_{SZ}, PS_{EA}). SVs and CNVs, variant types that cannot be ascertained comparably in exome and WGS datasets, were not included. To minimize ancestry as a confounder in PSs, analysis was restricted to a subset of 7,181 families ($n = 25{,}391$ individuals) with parents and offspring who have confirmed European ancestry.

The contribution of each factor individually and the additive contributions of multiple factors were estimated by multivariable regression (Fig. 2a). The variance explained by individual genetic factors in the present study was consistent with previous studies. Polygenic risk explained 2% of the variance in case status in the combined sample (Supplementary Table 12), consistent with the ~2% of variance explained by polygenic risk in the most recent GWAS meta-analysis[3]. The combined contribution of rare variants was similar, also explaining 2% of the variance in case status (Fig. 2a). Our results indicate that rare variants and polygenic risk form two major components of the genetic architecture of ASD, and the additive effects of all factors combined could be quantified in a single model ($r^2 = 4\%$;, Fig. 2a and Supplementary Table 12). We applied the estimates of the multivariable regression to create composite

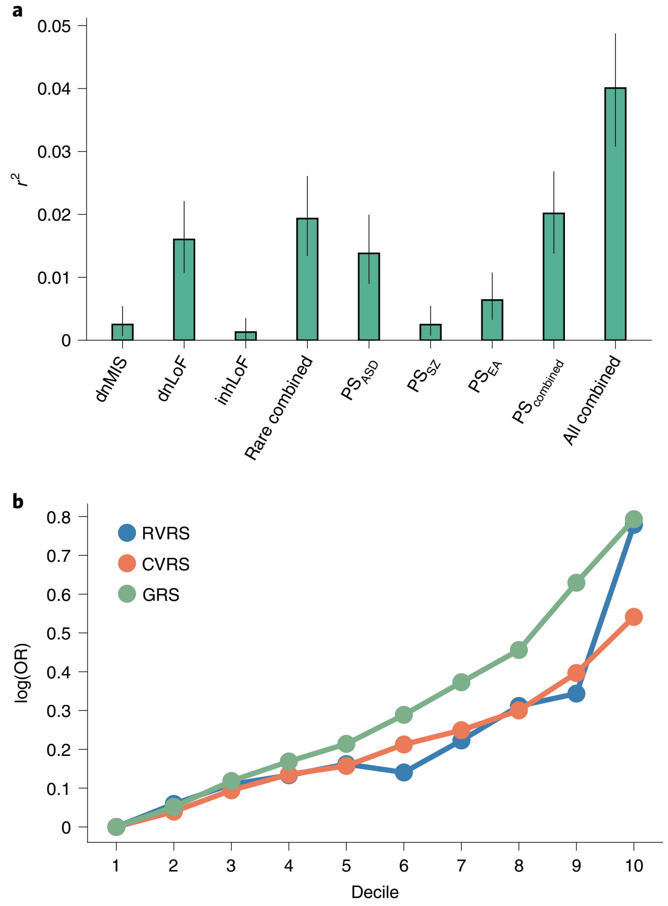


Fig. 2 | Multivariable regression of six genetic factors to create a composite GRS. **a**, Variance in case status explained (r^2 and 95% CI) by each genetic factor individually and in combination. Combined effects of rare variants (rare combined) PSs (PS combined) and all genetic factors (all combined) were estimated in the European ancestry sample ($n = 25,391$) by multivariable logistic regression controlling for sex, cohort and PCs. **b**, The $\log_{10}$(OR) of case-control proportions for the composite genetic risk scores RVRS, CVRS and GRS at multiple thresholds (deciles). Across all thresholds, effect sizes for the GRS were an average of 40% greater than for RVRS or CVRS alone. See the results in Supplementary Tables 13 and 14.

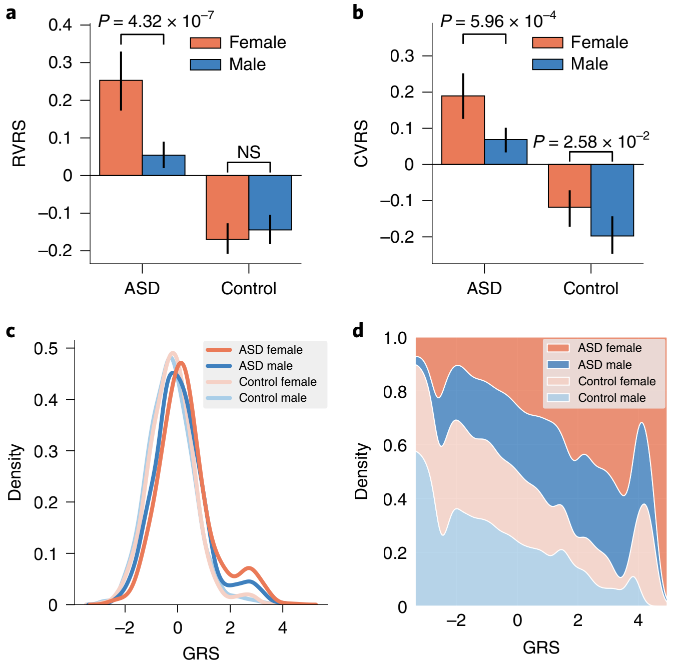


Fig. 3 | Increased genetic load in females with ASD compared with males. **a**,**b**, Increased burden of genetic risk in female cases compared with male cases is evident for the combined rare de novo and inherited variants (RVRS) (**a**) and combined PSs (CVRS) (**b**). NS, not significant. *P* values from a two-sided Student's *t*-test are shown. Participants consisted of 5,247 cases (4,256 males and 991 females) and 3,054 controls (1,504 males and 1,550 females) of European ancestry. **c**, Sex differences in the combined genetic load (GRS) are evident across the full distribution. **d**, A fill plot comparing the densities of distributions illustrating that the GRS of females (cases and controls) is skewed upward relative to males.

genetic risk scores of multiple factors, including a rare variant risk score (RVRS) for the combination of dnMIS, dnLoF and inhLoF, a common variant risk score (CVRS) for the combination of PS_{ASD}, PS_{SZ} and PS_{EA}, and a genomic risk score (GRS) for the combination of all six genetic factors. For each, we calculated the case–control ORs at multiple score thresholds (Fig. 2b and Supplementary Table 13) and found that, across the full distribution of risk scores, the GRS detects an effect size that is 40% stronger on average than effect sizes for RVRS or CVRS (Supplementary Table 14).

Sex differences in genetic load. Sex differences in genetic load were evident for both polygenic and rare variant risk (Fig. 3a,b). Female cases had significantly increased RVRS ($P = 4.32 \times 10^{-7}$; Fig. 3a) and CVRS ($P = 5.96 \times 10^{-4}$; Fig. 3b) compared with male cases. A similar trend was seen for polygenic risk in controls, with female controls having a greater CVRS than males ($P = 0.026$; Fig. 3b). These results are consistent with a 'female protective effect', in which females in the general population tolerate a greater genetic load of ASD risk, and likewise a greater genetic load is required for females to meet diagnostic criteria for ASD case status[21]. The full distribution of GRS is skewed upward in females compared with males (Fig. 3c), which is further highlighted by a fill plot comparing the densities of distributions of GRS between groups (Fig. 3d). As expected, the distribution of GRS is bimodal, with a subset of DNM carriers having the highest scores and the greatest enrichment of female cases.

According to a liability-threshold model for ASD[22,23], a total genetic load sufficient to meet diagnostic criteria can be reached through differing combinations of rare and common variation. Subjects with a greater rare variant load may require less polygenic load[5] and vice versa. In the present study, cases who carry damaging DNMs (dnLoF or dnMIS) had a combined polygenic load that was reduced compared with cases that do not carry damaging DNMs (Fig. 4). A similar trend was seen in both sexes, but the effect was not statistically significant in females. Thus, in the presence of a damaging DNM, less polygenic risk is required to meet diagnostic criteria for ASD. The negative correlation of the composite risk scores RVRS and CVRS ($P = 0.0037$, Pearson's correlation = −0.015) was stronger than for the pairwise correlations of individual factors (Fig. 4b and Supplementary Table 15), consistent with liability being attributable to the additive effects of multiple rare and common genetic factors. Also consistent with a liability threshold model, rare inherited variants (inhLoF) were negatively correlated with DNMs ($P = 0.03$; Extended Data Fig. 2a).

The strength of the threshold effect in Fig. 4b did not differ significantly by sex. This is in contrast to our previous analysis of this dataset using the polygenic scoring method PRSice, which found evidence that the anti-correlation of CVRS and RVRS was stronger in males[24] than in females (Extended Data Fig. 3). Evidence for sex differences in the strength of this negative correlation is therefore not robust across multiple polygenic scoring methods. Evidence for

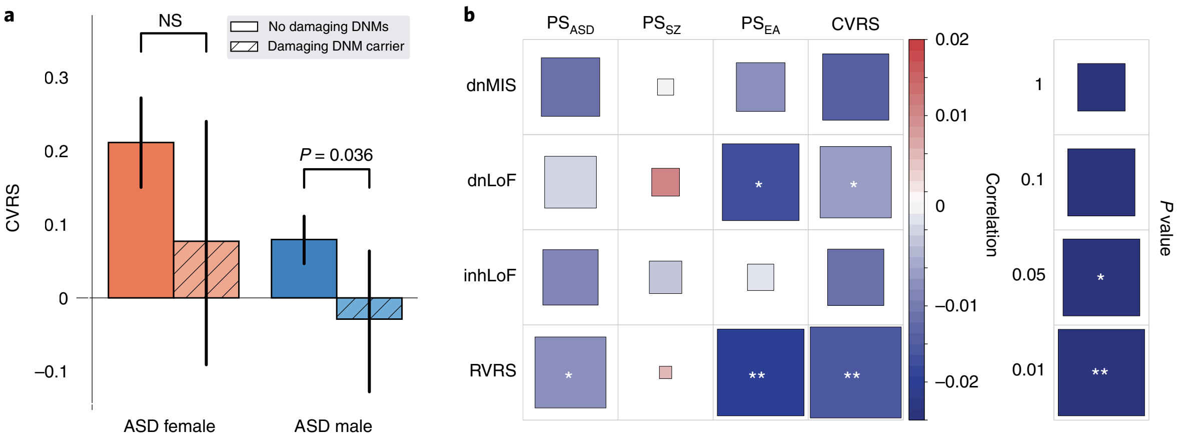


Fig. 4 | Negative correlation of rare variants and polygenic risk is consistent with a liability threshold model. a, Transmission of polygenic risk (pTDT) was reduced in cases that carry damaging DNMs (dnLoF and dnMIS combined), with a non-significant result in females. *P* values were based on two-sided Student' *t*-tests ($n = 4,256$ male cases (423 DNMs and 3,833 no DNMs) and 991 females (1,504 DNM and 1,550 no DNM) of European ancestry). **b**, A heatmap displaying the strength of the correlations between PSs and rare variants. *P* values were derived from linear regression. Results are provided in Supplementary Table 15.

sex-biased transmission of rare inhLoF variants within families was similarly weak. For instance, we did not observe a biased transmission of risk from the more 'protected' parent (mothers) to the more susceptible offspring (male cases; Extended Data Fig. 2b), as we have previously hypothesized[8]. Thus, we do not find evidence that gene-by-sex effects result in dramatic biases in the transmission of risk from parent to child (see Supplementary Note for additional discussion).

Differential effects of genetic factors on behavioral traits. We hypothesize that the differences in genetic architecture that we observed in the present study could underlie broad variation in clinical phenotype across the autism spectrum. DNMs have been associated with a more severe clinical presentation of ASD characterized by greater intellectual impairment[2,25] and delays in meeting developmental milestones[26,27]. PSs for cognitive traits have been associated with a clinical subtype of high-functioning 'Asperger' cases[18]. We investigated behavioral correlates of genetic factors in quantitative phenotype data from cases, sibling controls and parents that were available in the SSC and SPARK cohorts. Phenotypic measures in offspring included repetitive behavior (Repetitive Behavior Scale (RBS)), social responsiveness (Social Responsiveness Scale (SRS)), social communication (Social Communication Questionnaire (SCQ)), adaptive behavior (Vineland Adaptive Behavior Scale (VABS)) and developmental motor coordination (Developmental Coordination Disorder Questionnaire (DCDQ)). Behavioral traits in parents included ASD symptoms (SRS, Broad Autism Phenotype Questionnaire (BAPQ)), educational attainment (EA) and parental age at birth of the proband (Supplementary Table 16). Genetic effects were tested by linear regression controlling for cohort, age, sex and PCs, and the effects were also tested for gene-by-sex interactions (Supplementary Table 17).

Multiple genetic risk factors contributed to dimensions of ASD symptom severity in cases and in their typically developing sibling and parents. Six gene–trait correlations were significant after Bonferroni correction for 72 tests (Fig. 5a) and 18 showed nominal associations ($P \leq 0.05$). Social deficits (SCQ, SRS) in offspring were associated with polygenic risk (PS$_{ASD}$) and dnLoFs, and the same factors influenced social behavior (SRS, BAPQ) in parents (Fig. 5b), with PS$_{ASD}$ associated with social deficits and dnLoFs correlated with reduced symptom severity in parents consistent with a de novo etiology. Deficits in the VABS in offspring were weakly correlated with dnLoFs and polygenic risk (PS$_{ASD}$, PS$_{SZ}$). Deficits in motor coordination (the DCDQ) were associated with rare variants (dnMISs, dnLoFs, inhLoFs) but not with PSs. PS$_{EA}$ was protective for core ASD symptoms of repetitive behavior and social communication deficits in offspring and was also associated with reduced symptom severity in parents (BAPQ, EA). Intriguingly, multiple inherited genetic factors in parents (inhLoFs, PS$_{EA}$ and PS$_{SZ}$) were associated with parental age.

The correlations of genetic factors with behavioral traits were weakly sex biased. Eleven gene–trait relationships showed nominal evidence for an interaction by sex, but none was statistically significant after correction for multiple testing. These results suggest that the effects of most genetic factors on behavioral traits were similar in females and males. Among the weak interactions that were observed, most gene-by-sex effects (8/11) were observed in controls or parents. This may be attributable to either a reduced power to detect sex differences in case samples that are predominantly male or the homogenizing effects of clinical ascertainment of ASD cases. Sex differences in genetic effects were not exclusively male biased (5/11 had stronger effects in females). For example, genetic effects on SCQ in cases included two factors that were male biased (PS$_{ASD}$ and PS$_{SZ}$) and two that were female biased (inhLoF and PS$_{EA}$) (Fig. 5a). Perhaps the most striking example of gene-by-sex interaction was that all six factors showed evidence for differential effects on maternal and paternal age (Fig. 5b).

Multiple genetic factors contribute to parental age effects. We and others have demonstrated that advanced paternal age correlates with increased rates of germline mutation in offspring[28–30], consistent with parental age effects being attributable in part to DNMs that accumulate in the paternal germline. An alternative model by Gratten et al. has postulated that advanced paternal age could itself be a trait that is directly influenced by a genetic liability for ASD carried by the father[31]. A recent study has found evidence that PS$_{ASD}$ is positively correlated with paternal age[32], providing support for Gratten et al.'s model.

Our results demonstrate that the genetic basis of the parental age effect in ASD is highly multifactorial with contributions from DNMs, rare inherited variants and polygenic risk. For example, common (PS$_{EA}$) and rare (inhLoF) variation in fathers were associated with older and younger paternal age, respectively (Fig. 6a), and the correlation of PS$_{EA}$ with advanced parental age was even

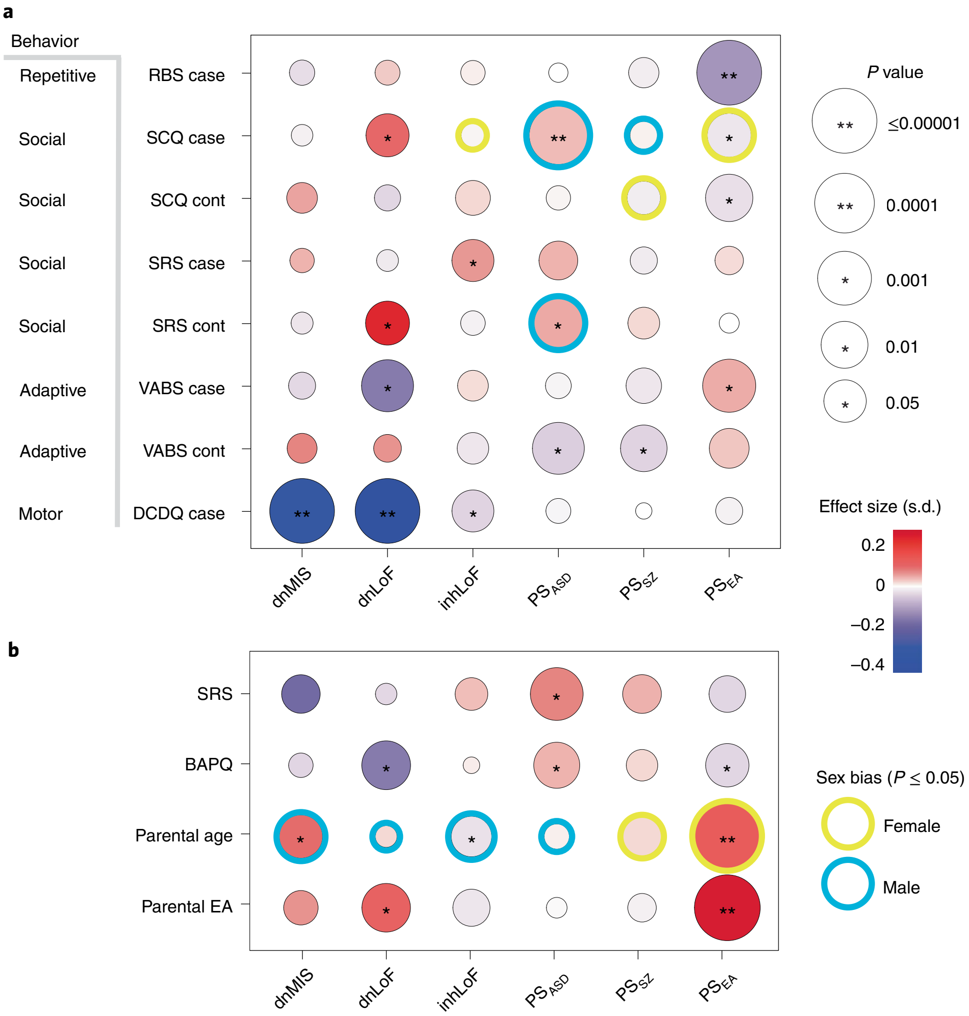


Fig. 5 | Differential effects of rare and common variation on behavioral traits in cases, sibling controls and parents. a, The effects of genetic factors tested on five phenotype measures in children: RBS, SRS, SCQ, VABS and DCDQ. Note that RBS, SRS, SCQ and BAPQ are measures of 'deficit'; thus, in the heatmap, red corresponds to increased severity. VABS and DCDQ are measures of 'skill'; thus, blue corresponds to increased severity on these two instruments. Gene–phenotype correlations were tested by linear regression controlling for sex, age, cohort and PCs. Effect size is given as the s.d. of the phenotype per unit of genetic factor. **b**, Genetic effects on parental behavior tested for autism-related symptoms (BAPQ, SRS), EA and parental age. In total, six gene–trait correlations were significant after Bonferroni correction for 72 tests ($^{**}P \leq 0.0007$), 18 were nominally significant ($^{*}P \leq 0.05$) and 11 showed evidence of sex-biased effects (gene-by-sex interaction, $P \leq 0.05$). Male or female sex bias indicates which sex had the greatest absolute value of effect size. Sample sizes for each phenotype ranged from 3,429 to 11,485. Sample numbers and results are summarized in Supplementary Tables 16 and 17. Analysis was restricted to individuals of European ancestry.

stronger for mothers (Fig. 6a and Supplementary Table 18). As expected, the rate of de novo SNVs increased with paternal age in the combined dataset ($r^2_{\text{paternal}} = 0.42$, $r^2_{\text{maternal}} = 0.27$; Extended Data Fig. 4), and dnLoF and dnMIS variants mirror this effect (Fig. 6a).

The single strongest inherited factor that influenced parental age was PS_{EA} ($r^2 = 0.017$; Supplementary Table 17), whereas PS_{ASD} and PS_{SZ} showed much weaker correlations ($r^2 \leq 0.0006$). Consistent with these results, parents' levels of education were significantly correlated with parental age in our sample and maternally biased ($r^2_{\text{maternal}} = 0.066$, $r^2_{\text{paternal}} = 0.023$), but social deficits in parents were not correlated with parental age (Supplementary Table 19). To further examine what behavioral traits in parents may explain inherited mechanisms of parental age effects, we compared the relative effects of genetic factors on the age, education and social behavior of parents (Fig. 6b,c). The effects of six genetic factors on parental age were positively correlated with their effect sizes for EA of parents (Pearson's correlation coefficient (PCC) = +0.76, $P = 0.0039$; Fig. 6b) and negatively correlated with their effect sizes for social deficits (PCC = −0.66, $P = 0.016$; Fig. 6c). These results suggest that inherited mechanisms of parental age effects on ASD risk in offspring may be driven by genetic effects on learning and education in parents rather than by effects on parental social behavior.

Rare variant risk is enriched in neurons of the fetal cortex. ASD susceptibility genes are preferentially expressed in the developing brain[18,27]. We hypothesize that differences in effect sizes and associated phenotypes between common variants and rare variants may be attributable, in part, to differences in the brain expression of their respective genes. In the present study, we confirmed that ASD susceptibility genes are enriched in fetal cortex and cortical cell types, and compared the degree of enrichment between protein-coding genes implicated by rare variants or by GWASs.

We applied a rare variant transmission and de novo association (TADA) test[33] to the combined data in the present study to define a set of 125 ASD susceptibility genes (TADA genes) with a false discovery rate (FDR) < 0.05, and we obtained a set of 114 high-confidence,

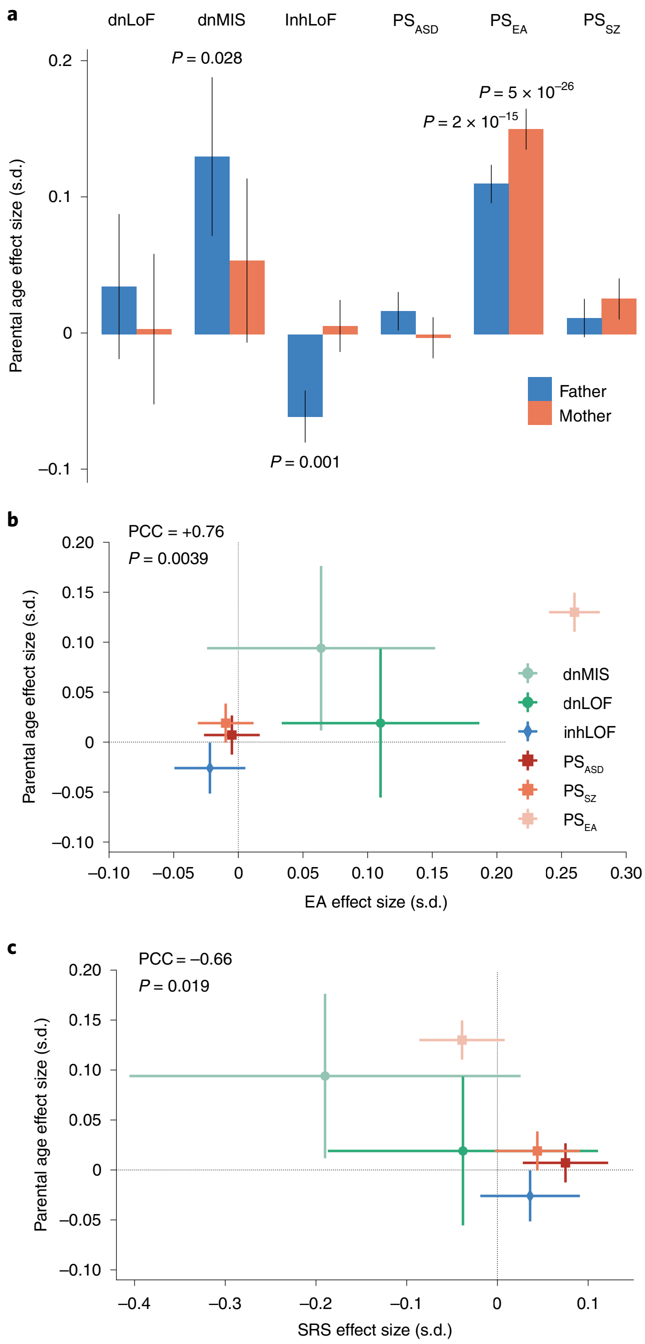


Fig. 6 | The genetic basis of parental age effects on ASD risk in offspring is multifactorial. a, Multiple genetic risk factors for ASD correlated with parental age with effects that differ by sex. Correlations of genetic factors with parental age (s.d. of age per unit of genetic load) were estimated for 11,485 individuals (5,749 mothers and 5,736 fathers). *P* values based on linear regression are given for individual effects with $P < 0.05$. Sex-stratified results for genetic effects on parental age are shown in Supplementary Table 18. **b**, The effects of six genetic factors on parental age positively correlated with their effects on EA and the strongest correlate of parental age was PS$_{EA}$. PCC, pearson correlation coefficient. **c**, The effects of six genetic factors on parental age negatively correlated with their effects on the SRS in parents. *P* values were derived from linear regression. Whiskers represent 95% CIs.

protein-coding genes identified in a previous GWAS by Grove et al.[18] (GWAS genes). To define a null distribution of expression values across developmental periods and cell types, 1,000 protein-coding genes were randomly sampled from the expression datasets. The three gene lists are provided in Supplementary Table 20. The expression of TADA genes and GWAS genes was then compared with the null distribution in cortex bulk tissue data from the BrainSpan transcriptome atlas[34] and cell-type expression data obtained from the Cortical development expression (CoDEx) resource[35].

In bulk human cortex, GWAS genes were more highly expressed (expression across all cortex samples and periods) compared with the null distribution, and TADA genes were further enriched (Fig. 7a). After normalizing cortex expression of each gene across periods, GWAS genes show increased relative expression during fetal development (Fig. 7b) compared with the null, and again TADA genes showed a further enrichment in fetal cortex. At the level of cortical cell types, the expression of TADA genes was significantly (approximately twofold) greater than the null in excitatory and inhibitory neurons (Fig. 7c and Supplementary Table 21), and GWAS genes did not show a significant enrichment of expression by cell type. These results are consistent with rare variants of large effect impacting genes that have key roles in early fetal brain development.

Discussion

Whole-genome analysis of a large ASD family cohort demonstrates how the genetic basis of ASD consists of multiple genetic components, including DNMs, rare inherited variants and PSs for psychiatric and behavioral traits. In the present study, the predictive accuracies of PSs and rare variants were similar, each explaining 2% of variance in case status. As new sequencing technologies continue to chip away at the missing heritability of ASD, additional genetic factors could be incorporated into the composite GRS to further improve upon this simple model. Furthermore, when WGS sample sizes become larger, more accurate estimates of the heritability explained by rare and common variants[36] could be feasible.

The genetic architectures of ASD vary across cases, which is evident by an inverse correlation of rare variants and PSs, consistent with a liability threshold model. This suggests that the genetic architectures of cases represent a spectrum of genetic loadings that span between extremes of polygenicity and monogenic disease. Furthermore, female cases have a significantly greater overall genetic load of polygenic and rare variation than male cases, confirming that a 'female protective effect', in which females display a greater tolerance for ASD risk alleles, applies generally to all components of the genetic architecture.

The spectrum of genetic architectures that we observe contributes to phenotypic variation across the cohort. Multiple genetic factors influence ASD symptom severity in cases and in their typically developing siblings and parents, with each factor having a different pattern of trait association. Considering core symptom domains such as social deficits and repetitive behavior, PS$_{ASD}$ and dnLoF were associated with severity in social deficits and PS$_{EA}$ was protective for these traits. Several factors were weakly correlated with adaptive behavior and deficits in developmental motor coordination were attributable solely to rare variants. For most gene–trait relationships, genetic effects on symptom severity paralleled their effects on case status. The one exception was PS$_{EA}$, which negatively correlated with symptom severity in offspring and parents but positively correlated with case status. Thus, the association of PS$_{EA}$ with ASD could not be explained by any of the behavioral traits that were measured in the present study. Potentially, SNPs that are captured by PS$_{EA}$ may influence dimensions of social cognition that were not tested in the present study or they may contribute to a clinically distinct subtype of high-functioning ASD. Consistent with the latter hypothesis, Grove et al. reported that the effect size for PS$_{EA}$ was strongest in the 'Asperger's syndrome' clinical subtype[18].

Based on the evidence for a 'female protective effect' on the genetic load in cases, one might predict that genetic effects on social

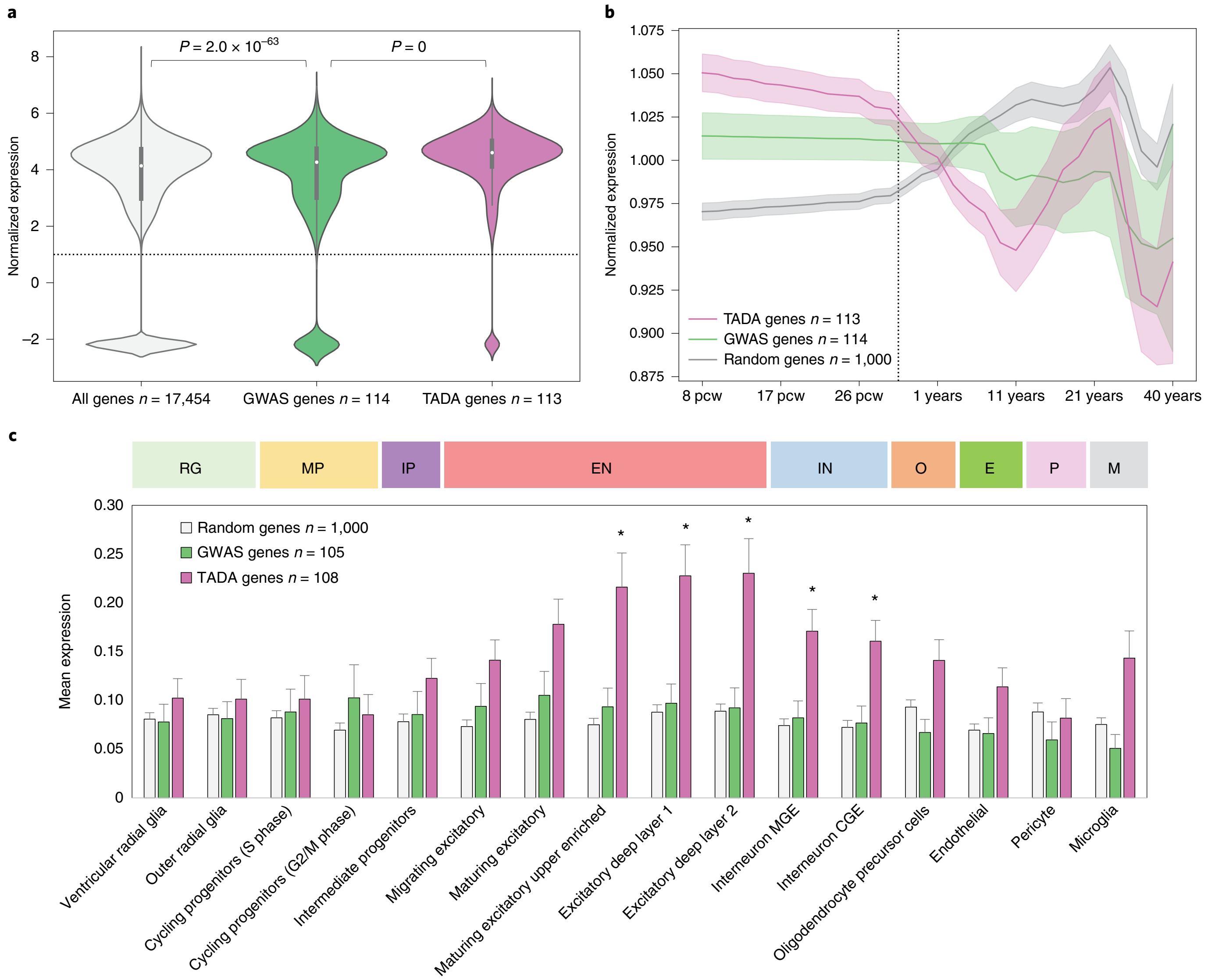


Fig. 7 | ASD susceptibility genes implicated by rare variants are enriched in neuronal cell types of the developing brain. Expression levels of protein-coding genes in bulk tissue (BrainSpan) and 16 cortical cell types (CoDEx) were compared across 115 genes identified with a rare variant association test in the present study (TADA) and 114 genes implicated by common variants in the study of Grove et al.[18] (GWAS). **a**, Expression of GWAS genes across all periods and brain regions were enriched relative to the full distribution and the expression of TADA genes were further enriched relative to GWAS genes. Boxes and whiskers represent the interquartile range (IQR) and 1.5× IQR, respectively. **b**, The expression of ASD genes in the developing cortex (after normalizing genes in BrainSpan to a mean expression of 1 across periods) enriched during prenatal development relative to null distribution consisting of 1,000 randomly protein-coding genes, with TADA genes being enriched to a greater extent. Shaded regions represent the mean of the 95% CIs from Lowess smoothing. **c**, Mean expression of the GWAS and TADA genes estimated within 16 cell types in the CoDEx dataset and compared with the null distribution of randomly sampled genes (811/1,000 genes that were included in CoDEx) using a two-sided Student's *t*-test. After Bonferroni's correction for 32 tests ($^{*}P \leq 0.0016$), expression of TADA genes was significantly increased relative to the null in five neuronal cell types. Error bars represent the s.e.m. and *P* values were derived from the two-sided Student's *t*-test. Gene sets and cell-type expression results are provided in Supplementary Tables 20 and 21. CGE, caudal ganglionic eminence; E, endothelial cell; EN, excitatory neuron; IN, interneuron; IP, intermediate progenitor; M, microglia; MGE, medial ganglionic eminence; MP, mitotic progenitor; O, oligodendrocyte precursor; P, pericyte; RG, radial glia.

behavior would be stronger in males than in females. However, gene–trait relationships did not consistently follow this pattern. Most gene–trait correlations did not differ by sex. Genetic effects on social communication in cases consisted of two factors with evidence of a male bias (PS_{ASD} and PS_{SZ}) and two with evidence of a female bias (inhLoF and PS_{EA}). Genetic correlations with parental age consisted of four factors that were paternally biased (dnMIS, dnLoF, inhLoF and PS_{ASD}) and two that were maternally biased (PS_{EA} and PS_{SZ}). The observation that gene-by-sex effects go both ways is consistent with studies that have found preliminary evidence that some ASD genes are prevalent in female cases and others are prevalent in males[37]. Caution is warranted when interpreting gene-by-sex interactions. Given that all ASD GWASs have included case samples that were predominantly male, PS_{ASD} may be over-represented in male-biased SNP effects. In addition, genetic effects that differ by sex could reflect the influences of social factors or clinical ascertainment[38]. For example, a female bias in the effects of inhLoF variants might be expected if the clinical ascertainment of females is biased toward subjects with greater symptom severity and greater rare variant load[39].

Multiple genetic factors were associated with parental age with effects that differed by sex. These results provide new insights into the genetic mechanisms of parental age effects on ASD risk in offspring[40]. Parental age effects are attributable to multiple mechanisms, including: (1) a DNM mechanism (dnMIS, dnLoF) in which new mutations accumulate with age in the paternal germline[28,29]; (2) inherited rare variants that directly contribute to parental age behavior in fathers; and (3) a polygenic mechanism that influences parental age in mothers and fathers[31], with PS_{EA} having by far the strongest effect. Our genetic findings support a model in which the combined effects of inhLoF, DNMs and PSs contribute to a U-shaped effect of parental age and genetic risk for ASD. This model is consistent with several previous studies that have found evidence for a U-shaped relationship of parental age and risk for ASD or other developmental disorders in offspring[41–45].

The effects of genetic factors on parental age were positively correlated with their effects on EA. Rare inhLoF variants were associated with early paternal age and fathers that carried inhLoFs had reduced EA, but this association was not statistically significant ($P<0.058$; Supplementary Table 18). The single strongest predictor of advanced parental age, particularly for mothers, was PS_{EA}. We confirmed in our dataset a significant correlation of parental education and parental age[46] that was stronger for mothers ($r^2=0.06$, $P=3.5\times10^{-52}$; Supplementary Table 19) than for fathers ($r^2=0.03$, $P=1.3\times10^{-23}$). By contrast, measures of social impairment in parents (SRS, BAPQ) were not associated with advanced parental age. Our results support a hypothesis that inherited mechanisms of parental age effects are mediated by genetic effects on learning and education in parents.

Differences in cognitive traits associated with rare variants and polygenic risk may be in part attributable to expression patterns of the respective genes during fetal development. By comparing the expression of GWAS and TADA genes in transcriptome data from bulk tissue and single cells of the developing cortex, genes implicated by rare variants were more strongly enriched during fetal development, specifically within neurons. These results are consistent with polygenic models in which rare variants impact genes that play key roles in neurodevelopment, whereas the effects of common risk alleles are distributed more broadly across genetic regulatory networks[47,48]. Given that much of the polygenic risk influences noncoding regulatory elements of genes[49], it is possible that the brain and cell-type enrichment of common variant effects may be greater for the underlying regulatory elements than for the transcripts as a whole. However, these results do highlight one aspect of the genetic architecture: polygenic risk for ASD is not restricted to a narrowly defined brain region, cell type or pathway.

The results described in the present study highlight how an integrated analysis of multiple genetic factors can improve our understanding of the genetic basis of ASD. Although most of the heritability of ASD remains unexplained, the expanding arsenal of sequencing platforms and methods of variant detection promise to expand the range of genetic factors that can be captured from a genome. The growing cohorts of ASD[19] as well as individual rare diseases[50] promise to improve knowledge of the effects of risk alleles on psychiatric traits and how their combined effects determine clinical outcome.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01064-5.

Received: 25 February 2021; Accepted: 28 March 2022;
Published online: 2 June 2022

References

1. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
2. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
3. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
4. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25**, 528–535 (2009).
5. Bergen, S. E. et al. Joint contributions of rare copy number variants and common SNPs to risk for schizophrenia. *Am. J. Psychiatry* **176**, 29–35 (2019).
6. Davies, R. W. et al. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat. Med.* **26**, 1912–1918 (2020).
7. Lim, E. T. et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235–242 (2013).
8. Zhao, X. et al. A unified genetic theory for sporadic and inherited autism. *Proc. Natl Acad. Sci. USA* **104**, 12831–12836 (2007).
9. Werling, D. M. & Geschwind, D. H. Recurrence rates provide evidence for sex-differential, familial genetic liability for autism spectrum disorders in multiplex families and twins. *Mol. Autism* **6**, 27 (2015).
10. Robinson, E. B., Lichtenstein, P., Anckarsater, H., Happe, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl Acad. Sci. USA* **110**, 5258–5262 (2013).
11. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
12. Desachy, G. et al. Increased female autosomal burden of rare copy number variants in human populations and in autism families. *Mol. Psychiatry* **20**, 170–175 (2015).
13. Jacquemont, S. et al. A higher mutational burden in females supports a 'female protective model' in neurodevelopmental disorders. *Am. J. Hum. Genet.* **94**, 415–425 (2014).
14. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
15. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
16. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
17. Weiner, D. J. et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).
18. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
19. Spark Consortium. SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).
20. Lloyd-Jones, L. R. et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
21. Werling, D. M. The role of sex-differential biology in risk for autism spectrum disorder. *Biol. Sex. Differ.* **7**, 58 (2016).
22. Falconer, D. S. Inheritance of liability to certain diseases estimated from incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
23. Reich, T., Morris, C. A. & James, J. W. Use of multiple thresholds in determining mode of transmission of semi-continuous traits. *Ann. Hum. Genet.* **36**, 163 (1972).
24. Antaki, D. et al. A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Preprint at *medRxiv* https://doi.org/2021.03.30.21254657 (2021).
25. Robinson, E. B. et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nat. Genet.* **48**, 552–555 (2016).
26. Buja, A. et al. Damaging de novo mutations diminish motor skills in children on the autism spectrum. *Proc. Natl Acad. Sci. USA* **115**, E1859–E1866 (2018).
27. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
28. Michaelson, J. J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).
29. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
30. Goriely, A. & Wilkie, A. O. Missing heritability: paternal age effect mutations and selfish spermatogonia. *Nat. Rev. Genet.* **11**, 589 (2010).
31. Gratten, J. et al. Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nat. Genet.* **48**, 718–724 (2016).
32. Mullins, N. et al. Reproductive fitness and genetic risk of psychiatric disorders in the general population. *Nat. Commun.* **8**, 15833 (2017).
33. He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).

34. Li, M. et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, eaat7615 (2018).
35. Polioudakis, D. et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801.e8 (2019).
36. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
37. Turner, T. N. et al. Sex-based analysis of de novo variants in neurodevelopmental disorders. *Am. J. Hum. Genet.* **105**, 1274–1285 (2019).
38. Russell, G., Steer, C. & Golding, J. Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Soc. Psychiatry Psychiatr. Epidemiol.* **46**, 1283–1293 (2011).
39. Werling, D. M. & Geschwind, D. H. Sex differences in autism spectrum disorders. *Curr. Opin. Neurol.* **26**, 146–153 (2013).
40. D'Angelo, D. et al. Defining the effect of the 16p11.2 duplication on cognition, behavior, and medical comorbidities. *JAMA Psychiatry* **73**, 20–30 (2016).
41. Malaspina, D. et al. Paternal age and intelligence: implications for age-related genomic changes in male germ cells. *Psychiatr. Genet.* **15**, 117–125 (2005).
42. Lyall, K. et al. The association between parental age and autism-related outcomes in children at high familial risk for autism. *Autism Res.* **13**, 998–1010 (2020).
43. Frans, E. M. et al. Autism risk across generations: a population-based study of advancing grandpaternal and paternal age. *JAMA Psychiatry* **70**, 516–521 (2013).
44. Lampi, K. M. et al. Parental age and risk of autism spectrum disorders in a Finnish national birth cohort. *J. Autism Dev. Disord.* **43**, 2526–2535 (2013).
45. Lundstrom, S. et al. Trajectories leading to autism spectrum disorders are affected by paternal age: findings from two nationally representative twin studies. *J. Child Psychol. Psychiatry* **51**, 850–856 (2010).
46. Fulco, C. J., Henry, K. L., Rickard, K. M. & Yuma, P. J. Time-varying outcomes associated with maternal age at first birth. *J. Child Fam. Stud.* **29**, 1537–1547 (2020).
47. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
48. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
49. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
50. Jacquemont, S. et al. Genes to mental health (G2MH): a framework to map the mombined effects of rare and common variants on dimensions of cognition and psychopathology. *Am. J. Psychiatry* **179**, 189–203 (2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Our research complies with all relevant ethical regulations as approved by the institutional review board of the UCSD School of Medicine.

Datasets. The sample comprised three datasets, including WGS of cohorts from the REACH project at UCSD (https://sebatlab.org/reach-project) and the SSC, and a dataset of exomes and SNP genotyping from the SPARK study[19]. The combined sample of 11,313 ASD families consisted of a total of 37,375 individuals, including 12,270 cases, 5,190 typically developing siblings and 19,917 parents (Supplementary Tables 1 and 2). All categories of genetic risk to be evaluated in the present study were confirmed previously within smaller cohorts of this study (REACH or SSC). Thus, the combined sample provides improved power to determine the effect sizes for the same genetic factors. See Data availability and Code availability for details on data access.

Processing of DNA-sequence data. Each of the three datasets consisted of Illumina paired-end sequence data, which were processed by BWA alignment and variant calling using GATK best practices. Specific differences between datasets include library prep (PCR versus PCR free, WGS versus exome) and differences in software version. Details are provided in the sections below. Analysis was carried out with SNP, indel and SV calls mapped to GRCh38. Variant calls from the SSC and SPARK cohorts were generated from sequences aligned to GRCh38. Jointly called variant call formats (VCFs) from the REACH cohort were lifted over from GRCh37 to GRCh38 before annotation and analysis.

REACH cohort. WGS was performed on blood-derived genomic DNA as described in our previous publication[51]. Standard quality control steps were carried out to ensure proper relatedness and genetic sex concordance with the sample manifest. Sequencing reads were aligned to the GRCh37 reference genome using bwa-mem (v.0.7.12). Subsequent processing of the alignments followed GATK Best Practices guidelines including sorting, marking duplicate reads, indel realignment and base quality score recalibration.

To ensure functional equivalency with other cohorts in our dataset, we applied the same SNV/indel variant calling pipeline used on the SSC cohort (see Simons Simplex cohort below for details). We utilized GATK HaplotypeCaller (v.4.1) to first call SNVs and indels in individual samples. GRCh38 genomic VCFs (GVCFs) were then combined using CombineGVCF and jointly genotyped. Variant Quality Score Recalibration (VQSR) was then performed on the joint VCF. The VQSR model was trained with the parameter 'maxGaussians=8' for SNVs or 'maxGaussians=4' for indels. Variant scores were recalibrated with the truth sensitivity level of 99.8% for SNVs and 99.0% for indels. Sample-level filtering converted genotypes to noncalls ('./.') if the genotype quality (GQ) < 20 or the read depth (DP) < 10. Before proceeding with variant annotation, variants were lifted over from GRCh37 to GRCh38 with the GATK LiftoverVcf command.

Simons Simplex cohort. WGS was performed at the New York Genome Center (NYGC) on an Illumina HiSeq X10 sequencer using 150-bp paired-end reads to an average depth of 40×. Reads were aligned to the GRCh38 reference genome using bwa-mem with subsequent processing of alignments in line with GATK Best Practices for functional equivalence.

Jointly called VCFs containing SNVs and indel calls were provided by the NYGC (21 March 2019). Briefly, variant calling was performed using GATK (v.3.5). Variant discovery implemented HaplotypeCaller in GVCF mode. Variant quality scores were recalibrated by VQSR with the truth sensitivity level of 99.8% for SNVs and 99.0% for indels. Low-quality genotype calls were defined as GQ < 20 or DP < 10 and were converted to missing genotypes ('./.'). Only variants that had 'PASS' entries in the FILTER column were considered for analysis of inherited variants. Further details on the generation of the SSC SNV and indel joint calls can be found in the pdf accompanying the data release from the Simons Foundation. WGS SNP genotypes and GWAS-imputed genotypes were subsequently merged in PLINK 1.9 (ref. [52]) for generation of PCs and PSs.

SPARK cohort. The publicly available SPARK dataset consisted of SNP genotyping (Illumina global screening array GSA-24v1-0) and exomes (IDT xGen capture sequenced on the Illumina NovaSeq 6000 using 2/S4 flow cells). Imputation of SNP genotypes was performed using the RICOPILI pipeline (https://sites.google.com/a/broadinstitute.org/ricopili/imputation)[53].

Downstream processing of exome data was performed as follows: per-sample GVCFs were obtained from the SPARK September 2019 data release in which GVCFs had been generated with GATK v.4.1.2.0 HaplotypeCaller from CRAM files aligned to GRCh38 with bwa-mem. Joint genotyping and quality control of SNP and indel variant calls were performed at UCSD in batches of 100 families, the same GATK pipeline that was used for the REACH and SSC WGS. Variants with 'PASS' in the FILTER column were retained for analysis. Likewise, indel calls with quality of depth (QD) < 7.5 were omitted.

PC calculation. Genotype data were linkage disequilibrium (LD) pruned to a set of 100,370 unambiguous markers with a minor allele frequency (MAF) >5% in PLINK 1.9, using the --indep-pairwise command with a 200-variant window, shifting the window 100 variants at a time and pruning variants with $r^2 > 0.2$. KING v.2.2.4 (https://doi.org/10.1093/bioinformatics/btq559) was used to identify a set of unrelated individuals (first- and second-degree relatives removed). PCs were calculated in the unrelated individuals based on LD-pruned data using FlashPCA2 (https://doi.org/10.1093/bioinformatics/btx299) and related individuals were then projected on to the PCs.

PS calculation. PS_{SZ} was calculated based on current schizophrenia summary statistics from the psychiatric genomics consortium (https://www.med.unc.edu/pgc/download-results). PS_{ASD} was calculated from summary statistics in Grove et al.[18] after excluding the SSC dataset used in the present study. PS_{EA} was calculated from summary statistics of the recent GWAS meta-analysis of EA by Lee et al.[54].

Two polygenic scoring methods were evaluated, and the method with the greatest prediction accuracy for ASD case status was selected for all analyses described here. Before manuscript submission, PSs were calculated from summary statistics using the method PRSice (v.2.3.0)[55] and the results of this analysis posted to *MedRxiv*[24]. As recommended during peer review, PSs were recalculated using a newer method, SbayesR[20]. The recalculated PSs, particularly PS_{ASD}, had greater predictive value for case status (Extended Data Fig. 5) and overall results were highly consistent between both PS methods. A comparison of the two is discussed in further detail in the Supplementary Note. SBayesR PSs were used for all analyses presented in the present study.

SBayesR. PSs were calculated using SBayesR[20], a polygenic scoring method that provides an advantage over 'clumping and thresholding' methods such as the method PRSice[55]. SBayesR utilizes all SNPs, and SNP effect sizes are re-scaled using a Bayesian (multiple regression) posterior inference model. SBayesR was implemented according to default settings as described in the software tutorial https://cnsgenomics.com/software/gctb/#Tutorial using the Banded LD matrix provided: https://cnsgenomics.com/software/gctb/#LDmatrices. We used the --exclude-mhc argument, which excludes variants in the major histocompatibility complex (MHC). Polygenic risk scores were calculated from the SBayesR summary statistics using PLINK.

PRSice v.2.3.0 (https://doi.org/10.1093/gigascience/giz082). Only unambiguous variants with an MAF > 1% in the reference dataset were included. Variants were LD clumped over a 250-kb window with an r^2 value of 0.1. PSs were calculated at multiple *P* value thresholds (0.01–0.9) to determine the optimal threshold. The best fitting PS for each trait was selected based on a significance level of a TDT test carried out in autism cases (*P* value threshold = 0.1 for ASD and SZ and 0.05 for EA). The best fitting PS was carried forward for all subsequent statistical analyses.

SV calling. SV calls were produced only for the WGS datasets REACH and SSC. Our SV calling and filtering workflow have been described in detail in our previous publication[51]. Briefly, we ran ForestSV, LUMPY and Manta on each sample calling deletions and duplications. ForestSV mainly relies on coverage as a feature to call SVs, resulting in segmented calls for large events that span repetitive elements such as segmental duplications. As a result of this, we applied a stitching algorithm to ForestSV calls, combining calls of the same SV class if they were ≤10 kb apart. As a preliminary filter, we omitted any variant that overlapped more than two-thirds of the SV length to centromeres, telomeres, segmental duplications, regions with low mappability with 100-bp reads, antibody parts, T-cell receptors and other assembly gaps.

The resulting calls were genotyped using SV2 and SVTyper within each sample. SVs and genotypes were then collapsed for overlapping calls. The collapsing algorithm first prioritized the breakpoint confidence interval (CIs) if both the start and the end CIs provided by LUMPY and/or Manta overlapped. For ForestSV calls, the CI was defined as ±100 bp from the start and end positions. The consensus position determined for a collapsible cluster was determined by the SV position with the highest number of overlaps. In the case of a tie, the median position was recorded. This method allows for collapsing of common SVs that 'tile' across a region, which rarely occurs outside variable regions such as the human leukocyte antigen locus. The resulting calls were then subject to a further round of collapsing, this time reducing calls to a consensus position if they overlapped 80% reciprocally with each other. This method was applied recursively until no more calls could be collapsed. As for CI collapsing, the consensus position reported was the SV with the highest number of overlaps. Variant level genotype likelihood scores were generated with SV2 by pooling all features from REACH and SSC samples. If the SV2 variant score was not 'PASS' then the SVTyper or Manta genotypes were recorded, as previously described[51]. Samples without a genotype call were considered as missing ('./.').

DNM calling. DNMs were called using the synthDNM software[56]. SynthDNM is a random forest-based classifier that uses only a pedigree file (PED/FAM) and VCF files as input, and can be readily optimized for different technologies or variant calling pipelines. For WGS datasets (REACH and SSC), we used the default SynthDNM classifier (SSC1 GATK), which was trained on GATK variant calls from >30× Illumina WGS data. This default classifier had high accuracy (area under the curve = 0.997) for detecting a truth-set of orthogonally validated de novo SNVs and indels from SSC[56]. For the exome dataset (SPARK), we trained an additional four classifiers, one for each set of variant calls: DeepVariant, WeCall, SPARK GATK and SSC GATK. To maximize sensitivity while controlling for false positives, we retained DNMs if they were called by three out of the five classifiers. To further confirm the accuracy of SPARK DNM calls, we compared the de novo SNV and indel calls on the SPARK dataset with a set of validated DNMs that were confirmed by Sanger sequencing from a previous pilot study[57]. For SNVs, the recall rate for SNVs ranged from 92.6% to 98.2% ($n = 117$), whereas for indels the recall ranged from 98.6% to 100% ($n = 107$). For further details of the methodology and performance of SynthDNM, refer to our companion paper[56].

De novo SVs were defined as events with heterozygous genotypes in offspring and homozygous reference genotypes in parents. We considered only variants that passed the stringent 'DENOVO_FILTER' filter produced by SV2 (ref. [58]). We applied our standard filtering guidelines detailed below to omit variants present in regions known to produce spurious calls. We also supplemented our de novo calls with the de novo CNV calls generated from microarrays in SSC samples from Sanders et al.[11] because many of these calls are likely to be missed by paired-end SV callers. We then manually inspected the list of de novo SVs and stitched calls together if they were separated by segmental duplications >10 kb (the maximum stitching requirement for ForestSV calls detailed in the section below).

Variant annotation. Variant Effect Predictor v.97 along with transcript annotations from Gencode v.31 were used in annotation of SNVs and indels. Variants were flagged as 'LoF' as the functional consequence of one of the following: 'transcript_ablation', 'splice_acceptor_variant', 'splice_donor_variant', 'stop_gained', 'frameshift_variant', 'stop_loss', 'start_loss'. LoF variants exclusive to nonsense-mediated decay transcripts were omitted from subsequent analysis. SVs were annotated by overlap with exons and proximal CREs including 5′-UTRs, transcription start sites and fetal brain promoters. As the list of annotated proximal CREs was in GRCh37, we lifted over the GRCh38 SV calls to GRCh37 for all subsequent analyses.

We assigned gnomAD LOEUF scores (v.2.1.1) to each LoF variant and CRE-SV. If a variant overlapped more than one gene, as in the case for large SVs, we recorded the minimum (most constrained) LOEUF score to that variant. Constraint was quantified for missense variants using the MPC scores[59]. These scores are available for the GRCh37 build of the human genome and were transposed to GRCh38 for analysis. Missense variants without MPC scores due to updates to the reference genome were not used in subsequent analysis. The recommended cutoffs to enrich for the top tier of constraint (LOEUF < 0.37; MPC > 2) were applied to de novo and rare inherited LoF variants.

Association tests. *Selection of variant types to be tested.* For the present study, we sought to define several major categories of rare variant and common variant risk and to investigate their combined effects on ASD risk and behavioral traits. We settled on six categories (three rare and three common) that all have strong prior evidence for their contribution to ASD.

Rare variants. The major categories of rare variants that have been reproducibly associated with associated with ASD include: (1) de novo protein truncating/dnLoFs, a category in which genetic association is concentrated within genes that are LoF intolerant[60]; (2) dnMISs, a category in which genetic association is concentrated within genes that show missense constraint[59]; and (3) inhLoF variants in LoF-intolerant genes[16,51]. In our analysis of genetic association, we confirmed the association of SNPs, indels and SVs within the above categories (Fig. 1). Analysis of the interactions between factors and correlations of genetic factors with behavioral traits across multiple cohorts were restricted to SNP and indel variants that can be detected across cohorts with comparable sensitivity.

Polygenic scores. Across a series of studies, schizophrenia[61,62] and EA[62,63] have stood out as traits that are correlated with polygenic risk for ASD. PS_{SZ} and PS_{EA} may not be the psychiatric trait scores that are most highly correlated with PS_{ASD}, but they are among the most well-powered GWASs. For this reason, these PSs were selected for the first family-based study that applied a pTDT test (also used in the present study) to demonstrate an overtransmission of PS_{ASD}, PS_{SZ} and PS_{EA} to cases[17]. This established for us the proof of concept for their inclusion in the present study. Given the high intercorrelation of genetic risk for a variety of other psychiatric disorders and traits[61], a rationale could be made for examining several more PSs, but, given the wide variety of equally valid and highly correlated traits and PSs to choose from, we sought to err on the side of simplicity and included these three as the main polygenic factors of interest.

Definitions of variant types. All rare variant categories described below consisted of private variants in which the alt allele was present in only one family in the present study and had an allele frequency <1% in gnomAD (v.2.1.1). Target categories included only variants in functionally constrained genes as defined below.

The dnMIS variants were defined as all private de novo missense SNVs with MPC scores >2. The dnLoF variants were defined as variant calls that were predicted to result in loss of protein function (truncation of a protein) and included stop-gain, frameshift, splice site and exonic deletion in an LoF-intolerant gene defined as LOEUF < 0.37, following the recommendations from gnomAD. SVs that intersected more than one gene were assigned a minimum LOEUF score (corresponding to the most constrained gene). The category of de novo synonymous variants (dnSyns) included all private de novo synonymous SNVs.

The inhLoF variants were defined as private SNV or indel variants with a 'PASS' entry in the 'FILTER' column and we removed variants with ≥5% missing calls across the cohorts. For inhLoF variants in the SPARK dataset, we applied one additional filter removing indel variants with QD scores <7.5.

For dnSVs and inherited LoF SVs, we included only private exonic SVs >50 bp in length and CRE-SVs ≥2.5 kb that passed the 'DENOVO_FILTER' from the SV2 software, which is a stringent filter recommended for ultra-rare variants.

De novo association. The burden of damaging DNMs (dnLoF, dnMIS) was compared between cases and controls using a two-sided independent Student's t-test reporting the two-sided P values. Results are provided for the set of DNMs in the combined sample. In addition, to evaluate the consistency of DNM ascertainment across the REACH, SSC and SPARK cohorts, dnLoF, dnMIS and dnSyn variants in all cohorts were compared by restricting DNMs to a common set of exome targets that was used in a previous publication by Iossifov et al.[2]. The dnSyn variants did not differ significantly between cases and controls in the combined sample (Extended Data Fig. 1a). In the SPARK cohort we observed a 1.1-fold excess of synonymous variants in cases (OR = 1.1, $P = 0.02$). This trend could be attributable to other factors, including chance or true ASD-associated non-coding variants. No quality metrics that were tested were correlated with case status in the SPARK dataset including coverage, transition:transversion (Ti/Tv) ratio, ratio of heterozygous to homozygous genotypes (Extended Data Fig. 1b) and paternal age (Extended Data Fig. 1c). Thus, variables could not be identified that explain a subtle baseline difference in the dnSyn burden, which could be included as covariates in a regression model. However, this very subtle effect does not contribute to a bias in the combined sample and cannot explain the strong associations reported for other categories of DNM (Fig. 1).

Inherited rare variant association. The number of transmissions and nontransmissions from parent to offspring was obtained using plink's '--tdt poo' command (v.1.9). Pooling of transmission and nontransmission counts for the TDT was done using the pytdt python package (https://github.com/sebatlab/pytdt). This package takes as input a data table containing a unique variant ID and counts for transmissions and nontransmissions in fathers and mothers for both cases and controls. Pytdt performs the pooling or group-wise analysis of private LoF variants and CRE-SVs by summing the counts of transmissions and nontransmissions for all variants encompassing a group. The package also reports ORs, CIs and other statistics commonly used for TDT analysis. We also conditioned the TDT according to the damaging DNM burden in the offspring using a binomial test for statistical significance of transmission distortion of private variants to cases or controls separately. For a summary of the TDT results and a list of all the private variants tested in the analysis, see Supplementary Tables 7–9.

The pTDT. According to methods from Weiner et al.[17], trio-based association of PSs (PS_{ASD}, PS_{SZ}, PS_{EA}) with ASD was tested with the pTDT, which tests the significance of the deviation (dev) of the child PS from the average PS of the parents.

$$\text{pTDT} - \text{dev} = \text{child PS} - \text{midparent PS}.$$

The P value was then calculated using a one-sided Student's t-test of pTDT-dev (Fig. 1c) with a population mean of 0. Results of the pTDT are reported in Supplementary Table 10.

Calculating composite risk scores RVRS, CVRS and GRS. We used multivariable regression to capture the combined effects of multiple genetic factors on case status. For rare variant factors, the predictor variables in the model consisted of rare variant burden counts for dnMIS, dnLoF and inhLoF. For PSs PS_{ASD}, PS_{SZ} and PS_{EA}, the predictor variables consisted of the pTDT-dev values of the trios. To calculate a composite genetic risk score, each predictor variable was first residualized for PCs and sex. Then estimates were calculated from a generalized linear model as follows:

$$y \sim x1 + x2 + x3 + \text{PCs} + \text{sex}$$

where y is case status and $x1$, $x2$ and $x3$ are residualized predictor variables for three genetic factors. PCs for all regression models consisted of the first ten PCs from the principal component analysis (PCA). Then, the composite risk score (RS) is calculated using r as:

$$\text{RS} = \text{predict}(\text{model}, \text{type} = \text{“response”}).$$

Each RS was than standardized by z-transformation. Predictor variables ($x1$, $x2$, $x3$, and so on) for each risk score consisted of:

$$\text{Rare variant risk score (RVRS)} : \text{dnMIS} + \text{dnLoF} + \text{inhLoF}$$

$$\text{Common variant risk score (CVRS)} : \text{PS}_{\text{ASD}} + \text{PS}_{\text{SZ}} + \text{PS}_{\text{EA}}$$

$$\text{Genomic risk score (GRS)} : \text{dnMIS} + \text{dnLoF} + \text{inhLoF} + \text{PS}_{\text{ASD}} + \text{PS}_{\text{SZ}} + \text{PS}_{\text{EA}}.$$

To compare the effect sizes on case status for the genetic factors and the composite risk scores (Fig. 2b), Nagelkerke's r^2 values were calculated for each of the residualized predictor variables and for each composite risk score.

Pairwise correlations of rare variants and polygenic risk. To test the correlations between rare variants and polygenic risk, we constructed pairwise linear models:

$$y \sim x + \text{sex} + \text{cohort} + \text{case.status} + \text{PCs}$$

where the variable y is a polygenic score (PS_{ASD}, PS_{SZ}, PS_{EA} or CVRS) and x is a measure of rare variant load (dnLoF, dnMIS, inhLoF or RVRS). Gene-by-sex interaction was then tested in the following model:

$$y \sim x + \text{sex} + x^{*}\text{sex} + \text{cohort} + \text{case.status} + \text{PCs}.$$

Supplementary Table 15 contains the full results for all pairwise correlation of rare and polygenic risk conditioned on sex.

Effects of genetic factors on behavioral traits. The effects of genetic factors on behavioral traits were investigated in the SSC and SPARK cohorts using clinical phenotype data available from the Simons Foundation Autism Research Initiative (SFARI; see Data availability and Code availability). To eliminate confounders due to ancestry, only individuals of European ancestry confirmed by PCA were included. Clinical measures of ASD symptoms and related behaviors were selected that were available for cases, typically developing sibling or parents. Phenotype measures consisted of the summary scores from the DCDQ of motor function and the RBS that were available on cases, and the VABS, SCQ and SRS that were available on both cases and sibling controls. Behavioral phenotypes available on parents included the BAPQ, parental EA (from the background history questionnaire) and parental age at birth (for the children with ASD diagnosis). Phenotype measures that were available for both the SSC and the SPARK cohorts were normalized within cohort by z-transformation, then combined, and the cohort was included as a covariate in the downstream analyses. A summary of the sample sizes available for each phenotype measure is provided (Supplementary Table 16).

Association of genetic factors with developmental traits was tested by linear regression controlling for sex, cohort and PCs. In addition, a gene-by-sex interaction was tested to determine whether genetic effects on cognitive traits differed for males and females. Phenotypes in offspring (cases and siblings) were tested using the model:

$$y \sim x + \text{sex} + \text{age} + \text{cohort} + \text{PCs}$$

where y is the phenotype variable and x the genetic factor (DNMs, inhLoFs and PSs). In addition, a gene-by-sex interaction was then tested in this model:

$$y \sim x + \text{sex} + x^{*}\text{sex} + \text{age} + \text{cohort} + \text{PCs}.$$

Brain and cell-type expression of ASD susceptibility genes. The lists of TADA, GWAS and randomly selected protein-coding genes are provided in Supplementary Table 20. The expression of TADA genes and GWAS genes was compared in the developing human brain using the publicly available gene expression matrix from BrainSpan[34]. The two gene sets were also compared across 16 cell types in the human cortex using cell-type expression data available from the CoDEx dataset[35].

TADA genes. We defined a set of genes implicated by rare variants with the TADA[33] in our combined sample, using the recommended parameters for ASD relative risk and mutational rates for LoF and missense variants calculated by Samocha et al.[64]. TADA genes were defined as a set of 113 ASD genes that was associated with ASD at an FDR < 0.05.

GWAS genes. We obtained the list of high-confidence genes that were implicated by GWAS associations and described by Grove et al.[18] (GWAS genes). Briefly, genes that are probable contributors to GWAS associations were defined with H-MAGMA, a method that assigns noncoding SNPs to their genes based on long-range interactions detected by Hi-C in fetal and adult brain[65]. A list of 121 GWAS genes was provided by the authors (Hyejung Won, personal communication). To facilitate a valid comparison of genes implicated by rare variants and common variants, the GWAS gene set was restricted to a subset of 114 genes that were protein coding according to grch38 Ensembl gene annotations.

Random genes. Patterns of expression across developmental periods (Fig. 7b) and cell types (Fig. 7c) for GWAS genes and TADA genes were compared with null distributions obtained by randomly sampling 1,000 protein-coding genes from the BrainSpan and CoDEx datasets.

Analysis of gene expression in bulk tissue (BrainSpan). The developmental transcriptome dataset was downloaded from BrainSpan (https://www.brainspan.org/static/download.html), which consisted of normalized gene expression data from 26 brain structures (including 21 within the cortex) across 31 developmental time periods. Overall expression of GWAS and TADA genes in the developing cortex was compared by combining expression values across cortex samples and gene sets were compared with the null distribution by Student's t-test. Likewise, patterns of expression in cortex across developmental time periods were compared across gene sets by first normalizing the cortex expression of each gene to its mean across cortex samples, and then fitting the expression values of each gene set by Lowess smoothing using the 'Lowess' function described here: https://james-brennan.github.io/posts/lowess_conf.

Analysis of gene expression in 16 cell types from fetal cortex (CoDEx). Analysis was performed on cell-type gene expression values provided in the CoDEx dataset[35], which consisted of single-cell RNA-sequencing obtained by DropSeq analysis of sections of germinal zone and ventricular zone tissue from mid-gestation fetal cortex[66]. Briefly, in Polioudakis et al., raw counts were normalized and cells were clustered using Seurat (v.2.3.4)[67] and mean gene expression values per cell were calculated for genes in 16 cortical cell types. Cell-type expression values were obtained from the 'Genes' table on the CoDEx web interface (http://solo.bmap.ucla.edu/shiny/webapp) for TADA genes and GWAS genes, and these were compared with a random sampling of 1,000 protein-coding genes.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

WGS data from the SSC and exome and SNP genotyping data from SPARK are available from SFARI (https://www.sfari.org/resource/autism-cohorts). Summary genetic data from WGS and exomes, including individual counts for dnLoF, dnMIS, inhLoF and PSs for all subjects in the present study and input data files for all analysis code, are also available from SFARI. WGS data from the REACH project are available from the National Institute of Mental Health Data Archive (NDA), including the structural variant callset and raw sequence (FASTQ), alignment (BAM) and VCF files from the REACH cohort (https://nda.nih.gov/edit_collection.html?id=2019). GWAS summary statistics are available from the Psychiatric Genomics Consortium (ASD and SZ) (https://www.med.unc.edu/pgc/download-results) and the Social Science Genetic Association Consortium (EA) (http://www.thessgac.org/data). Bulk tissue expression data on ASD susceptibility genes was obtained from the BrainSpan developmental transcriptome dataset (v.0; https://www.brainspan.org/api/v2/well_known_file_download/267666525). Cell-type expression levels of ASD susceptibility genes in fetal cortex were obtained through the web interface of the CoDEx viewer (http://solo.bmap.ucla.edu/shiny/webapp).

Code availability

Analysis code for all major statistical genetic analyses in the paper and for generating Figs. 1–7 is available as a Google Colab notebook on Github (https://github.com/sebatlab/Antaki2021).

References

51. Brandler, W. M. et al. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
52. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
53. Lam, M. et al. RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2020).
54. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
55. Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
56. Lian, A., Guevara, J., Xia, K. & Sebat, J. Customized de novo mutation detection for any variant calling pipeline: SynthDNM. *Bioinformatics* **37**, 3640–3641 (2021).
57. Feliciano, P. et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *npj Genom. Med.* **4**, 19 (2019).
58. Antaki, D., Brandler, W. M. & Sebat, J. SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* **34**, 1774–1777 (2018).
59. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* (2017).
60. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
61. Brainstorm Consortium et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
62. Hagenaars, S. P. et al. Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N = 112 151) and 24 GWAS consortia. *Mol. Psychiatry* **21**, 1624–1632 (2016).
63. Clarke, T. K. et al. Common polygenic risk for autism spectrum disorder (ASD) is associated with cognitive ability in the general population. *Mol. Psychiatry* **21**, 419–425 (2016).

64. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
65. Sey, N. Y. A. et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020).
66. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
67. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

Acknowledgements

We thank the families who participated in the genetic studies of the REACH, SSC and SPARK cohorts. The large samples used in the present study were made possible through the strong commitment to sharing of individual-level data by SFARI, the National Institutes of Health (NIH) and UCSD, and the sharing of summary-level data by the Psychiatric Genomics Consortium. We thank W. Pfeiffer, the San Diego Supercomputer Center and Amazon Web Services for hosting the computing infrastructure necessary for completing this project. We thank W. Chung and SFARI for providing data and materials for validation of DNM calls. We thank N. Baya for assistance with data processing. We thank H. Wong for providing the list of high-confidence GWAS genes. This work was supported by grants to J.S. from the SFARI (grant no. SFARI 606768), NIH (grant nos. MH113715, MH119746, 1MH109501) and the Escher fund for Autism (grant no. 20171603), a grant to C.M.N. from the NIH (grant no. MH106595), and a grant to A.R.M. from the NIH (MH127077). L.M.I. was supported by grants from the NIH (MH109885, MH108528, MH105524, MH104766) and SFARI (#345469). D.A. was supported by a T32 training grant from the NIH (grant no. GM008666). M.K. was supported by grants from the Dutch Research Council (grant nos. 09150162010073 and 45219212).

Author contributions

J.S. conceived and coordinated the study. D.A., M.G. and J. Guevara performed data processing, variant calling and annotation of WGS and exome datasets. J.S., C.M.N. and A.X.M. supervised statistical genetic analyses. D.A., J. Guevara, A.X.M., M.K. and J.S. performed statistical genetic analyses. J. Guevara, D.A., L.M.I. and J.S. performed analysis of RNA-sequencing datasets. E.R., C.E.C. and J. Grove performed analysis of SNP genotypes and meta-analysis of summary statistics. J.S., C.C., K.K.V., A.H., M.J.A., K.P., E.C., J.G.G., A.R.M. and O.H. coordinated recruitment and DNA sample processing for the UCSD dataset.

Competing interests

A.R.M. is a co-founder and has an equity interest in TISMOO, a company focusing on applications of genetics and human brain organoids to personalized medicine. The terms of this arrangement have been reviewed and approved by the UCSD, in accordance with its conflict-of-interest policies. The remaining authors declare no competing interests.

Additional information

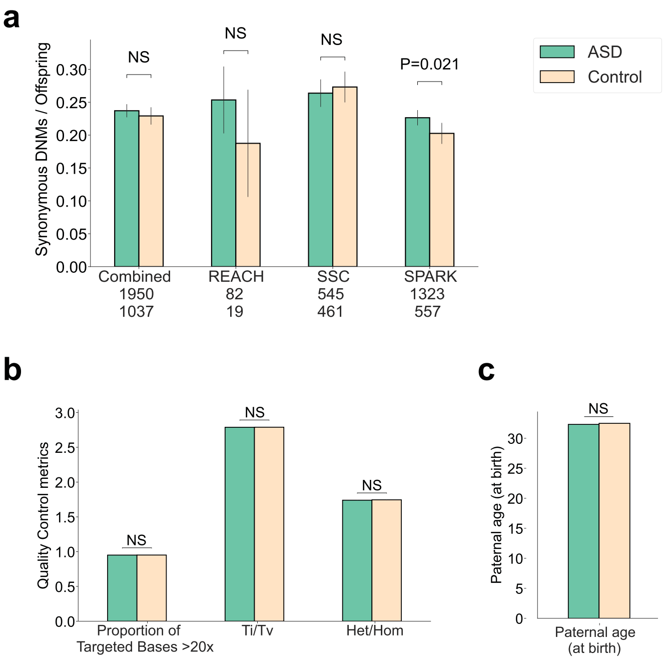
Extended data Extended data are available for this paper at https://doi.org/10.1038/s41588-022-01064-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01064-5.

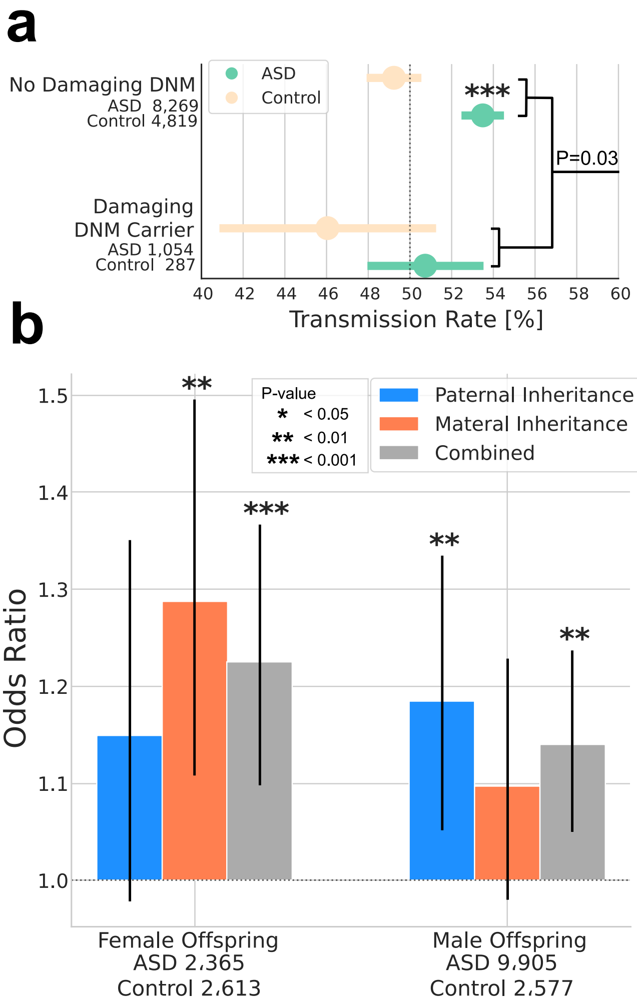
Correspondence and requests for materials should be addressed to Jonathan Sebat.

Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work.

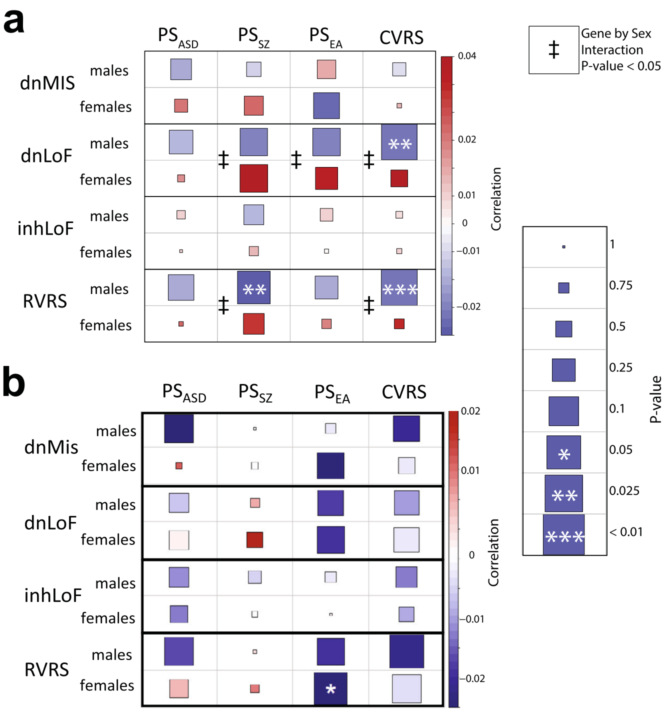
Reprints and permissions information is available at www.nature.com/reprints.



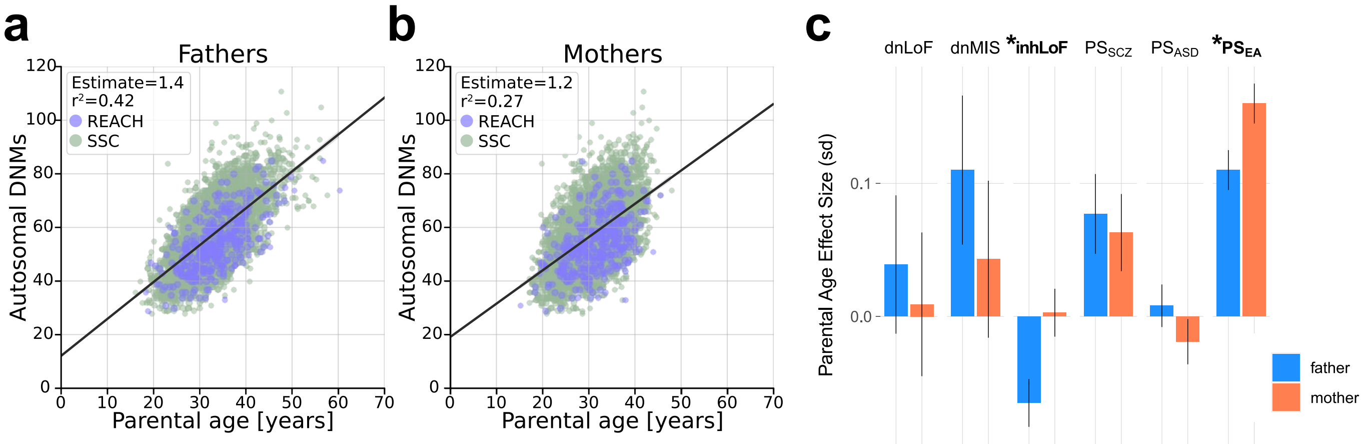
Extended Data Fig. 1 | Rates of de novo mutations stratified by cohort and evaluation of potential confounders. a, Rates of de novo synonymous (dnSyn) variants were not associated with ASD in the combined sample, but were enriched 1.1-fold in the SPARK cohort ($P=0.021$). **b**, We evaluated whether quality metrics or other confounders could explain the slight excess of dnSyn variants in SPARK cases. Quality metrics did not differ in cases and controls including coverage, transition:transversion ratio (Ti/Tv) or ratio of heterozygous calls (Het/Hom). **c**, Paternal age did not differ significantly between cases and controls.



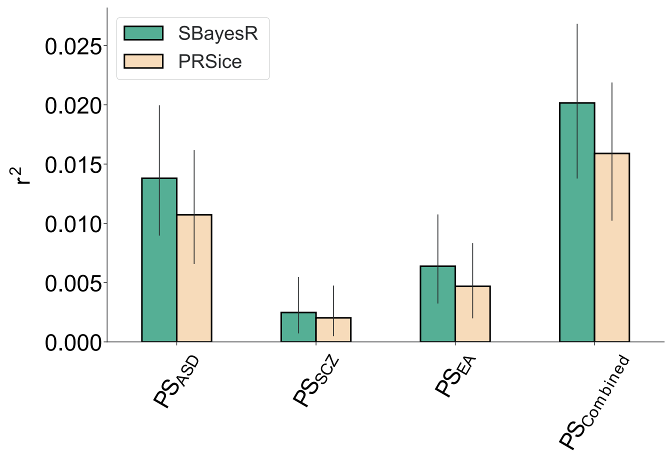
Extended Data Fig. 2 | The combined effects of dnLoF, inhLoF and sex on the transmission of rare variants in families. a, A significant liability threshold for rare variants was evident based on a negative correlation of dnLoF and inhLoF (linear regression *P*=0.03), and this effect did not differ significantly by sex. **b**, Case-control odds ratios were compared for the transmission rates in families by sex (father-daughter, mother-daughter, father-son, mother-son). Both maternal and paternal rare variants contribute to ASD with a significant over-transmission from mother to daughter and from father to son. We did not observe a significant sex bias in the transmission of rare variants in families. In particular, we did not observe an enriched transmission from mother to male cases as we have previously hypothesized[8].



Extended Data Fig. 3 | Sex differences in the correlation of rare variant and common variant risk was not robust across multiple polygenic scoring methods. a, An early analysis of this dataset using polygenic score estimates from PRSice observed that the negative correlation of RVRS and CVRS was stronger in males than in females, consistent with males having less tolerance of genetic risk. The heatmap displays the correlations between polygenic scores and rare variants in males and females separately. Correlations were tested by linear regression controlling for cohort, case status and ancestry PCs, and a gene-by-sex interaction was tested in the combined sample (‡gene-by-sex $P < 0.05$). **b**, With polygenic scores calculated using SBayesR, there was a similar trend with the correlation of CVRS and RVRS being stronger in males; however, the gene-by-sex interaction was not statistically significant.



Extended Data Fig. 4 | Correlation of de novo mutation rate with parental age. a,b, Correlation of total autosomal *de novo* SNVs with age of fathers (**a**) and mothers (**b**). See also Fig. 6a. *n* = 4,518 trios for which age-at-birth was available for the mother and father.



Extended Data Fig. 5 | Comparison of the predictive values of polygenic scoring methods PRSice and SBayesR. Polygenic scores calculated using SBayesR had greater predictive value for polygenic scores for ASD (PS_{ASD}), schizophrenia (PS_{SZ}) and educational attainment (PS_{EA}).

nature portfolio

Corresponding author(s): Jonathan Sebat

Last updated by author(s): Jan 25, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All publicly available software tools used for data processing are described in detail in the methods.
Data analysis	Analysis code for all major statistical genetic analyses in the paper and for generating all Figures 1 through 7 is available as a Google Colab notebook on Github (https://github.com/sebatlab/Antaki2021).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

WGS data from the SSC and Exome and SNP genotyping data from SPARK are available from the Simons Foundation Autism Research Initiative (SFARI) (https://www.sfari.org/resource/autism-cohorts). Summary genetic data from WGS and exomes including individual counts for dnLoF, dnMIS, inhLoF and polygenic scores for all subjects in this study and input data files for all analysis code are also available from SFARI. WGS data from the REACH project are available from the NIMH Data Archive (NDA) including the structural variant callset, and raw sequence (FASTQ), alignment (BAM) and variant call (VCF) files from the REACH cohort (https://nda.nih.gov/edit_collection.html?id=2019). GWAS summary statistics are available from the Psychiatric Genomics Consortium (ASD and SZ) (https://www.med.unc.edu/pgc/download-results/) and the Social Science Genetic Association Consortium (EA) (http://www.thessgac.org/data). Bulk tissue expression data on ASD susceptibility genes was obtained from the Brainspan developmental transcriptome dataset (v10; https://www.brainspan.org/api/v2/well_known_file_download/267666525). Cell type expression levels of ASD susceptibility genes in fetal cortex was obtained through the web interface of the Cortical development expression viewer (CoDEx) (http://solo.bmap.ucla.edu/shiny/webapp/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Updated sample size description: The combined sample of 11,313 ASD families consisted of a total 37,375 subjects including 12,270 cases, 5,190 TD siblings and 19,917 parents (supplementary tables 1-2). All categories of genetic risk to be evaluated in this study were confirmed previously within smaller cohorts of this study (REACH or SSC). Thus the combined sample provides improved power to determine effect sizes for the same genetic factors.
Data exclusions	It is well established that polygenic scores derived from sum stats of GWAS that are predominantly of European ancestry are informative ONLY in samples of european ancestry and are heavily confounded in other populations. In order to minimize the confounding effects of ancestry on polygenic scores (PS), association tests for PS and tests for the joint effects of rare variants and PS were performed on a subset of 7,181 families (N = 25,391 subjects) for which european ancestry was confirmed by principal component analysis.
Replication	See "selection of variant types to be tested". For this study we sought to define several major categories of rare variant and common variant risk and to investigate their combined effects on ASD risk and behavioral traits. We settled on six categories (3 rare and 3 common) that all have a strong prior evidence for their contribution to ASD. Rare variants. The major categories of rare variants that have been reproducibly associated with associated with ASD include (1) de novo protein truncating/loss-of-function mutations (dnLoF), a category where genetic association is concentrated within genes that are loss-of-function intolerant 60; (2) de novo missense variants (dnMIS), a category where genetic association is concentrated within genes that show missense constraint 59 and (3) inherited loss-of-function (inhLoF) variants in LOF-intolerant genes 16,52. In our analysis of genetic association we confirmed the association of SNPs, indels and SVs within the above categories (Fig. 1). Analysis of the interactions between factors and correlations of genetic factors with behavioral traits across multiple cohorts was restricted to SNP and indel variants that can be detected across cohorts with comparable sensitivity. Polygenic scores. Across a series of studies, schizophrenia 61,62 and educational attainment 62,63 have stood out as traits that are correlated with polygenic risk for ASD. PSSZ and PSEA may not be the psychiatric trait scores that are most highly correlated with PSASD, but they are among the most well powered GWASs. For this reason, these polygenic scores were selected for the first family-based study that applied a polygenic TDT test (also used in this study) to demonstrate an overtransmission of PSASD, PSSZ and PSEA to cases 17. This established for us the proof of concept for their inclusion in this study. Given the high intercorrelation of genetic risk for a variety of other psychiatric disorders and traits 61, a rationale could be made for examining several more polygenic scores, but given the wide variety of equally-valid and highly correlated traits and polygenic scores to choose from, we sought to err on the side of simplicity and included these three as the main polygenic factors of interest.
Randomization	This study followed a trio-based design. All cases included in this study were diagnosed with ASD using DSM criteria and standard diagnostic research instruments (ADI and ADOS). All controls consisted of parents and matched sibling controls. Major confounding factors were controlled for as covariates in statistical models including sex, cohort and ancestry principal components.
Blinding	All data and variant calls are publicly available and were obtained using automated pipelines designed to label specific genetic categories (de novo mutations, inherited LOF variants, polygenic scores) that have been defined in previous studies.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The cohort of 30K samples includes representatives from a wide variety of racial and ethnic backgrounds which are well known con founders in genetic studies. As such, all tests for genetic association consisted of trio-based tests that are well controlled for population stratification, all statistical models controlled for ancestry principal components, and analyses of polygenic scores derived from summary statistics of european cohorts were restricted to a subset of samples with confirmed european ancestry. Genetic association tests included standard sample covariates for cohort, sex, age and ancestry principal components
Recruitment	This study utilized archival dataset and did not prospectively recruit stubjects
Ethics oversight	The study was performed under a protocol that was approved by the IRB of the University of California San Diego

Note that full information on the approval of the study protocol must also be provided in the manuscript.