The female protective effect against autism spectrum disorder

Emilie M. Wigdor[1,2], Daniel J. Weiner[1,3], Jakob Grove[4-7], Jack M. Fu[1,9], Wesley K. Thompson[10], Caitlin E. Carey[1,3], Nikolas Baya[1,3], Celia van der Merwe[1,3], Raymond K. Walters[1,3], F. Kyle Satterstrom[1,3], Duncan S. Palmer[1,3], Anders Rosengren[7,8], Jonas Bybjerg-Grauholm[7], iPSYCH Consortium, David M. Hougaard[7], Preben Bo Mortensen[4,8,12,13], Mark J. Daly[1,3,11], Michael E. Talkowski[1,9], Stephan J. Sanders[14], Somer L. Bishop[14], Anders D. Børglum[4,5,8], Elise B. Robinson[1,3,15]*

1. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
2. Wellcome Trust Sanger Institute, Hinxton, UK
3. Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
4. Center for Genomics and Personalized Medicine (CGPM), Aarhus University, Aarhus, Denmark
5. Department of Biomedicine (Human Genetics) and iSEQ center, Aarhus University, Aarhus, Denmark
6. Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark
7. The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark
8. Institute of Biological Psychiatry, MHC Sct Hans, Copenhagen University Hospital, Copenhagen, Denmark
9. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA
10. Herbert Wertheim School of Public Health, University of California, San Diego, San Diego, California, USA
11. Finish Institute for Molecular Medicine, University of Helsinki, Helsinki, Finland
12. National Centre for Register-Based Research, Aarhus University, Aarhus, Denmark
13. Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark
14. Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, California, USA
15. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

*Correspondence to erob@broadinstitute.org

Abstract

Autism spectrum disorder (ASD) is diagnosed 3-4 times more frequently in males than in females. Genetic studies of rare variants support a female protective effect (FPE) against ASD. However, sex differences in common, inherited genetic risk for ASD are less studied. Leveraging the nationally representative Danish iPSYCH resource, we found siblings of female ASD cases had higher rates of ASD than siblings of male ASD cases ($P < 0.01$). In the Simons Simplex and SPARK collections, mothers of ASD cases carried more polygenic risk for ASD than fathers of ASD cases ($P = 7.0 \times 10^{-7}$). Male unaffected siblings under-inherited polygenic risk ($P = 0.03$); female unaffected siblings did not. Further, female ASD cases without a high-impact *de novo* variant over-inherited nearly three-fold the polygenic risk of male cases with a high-impact *de novo* ($P = 0.02$). Our findings support a FPE against ASD that includes common, inherited genetic variation.

Introduction

Autism spectrum disorder (ASD) is diagnosed three to four times more frequently in males than in females. The possibility of a 'female protective effect' (FPE) against ASD has been described extensively and has received consistent support from the results of genetic studies focusing on rare and *de novo* variants.[1–6] Many types of ASD-associated *de novo* variants are observed more frequently in female cases.[1–6] In general, the more ASD risk carried by a *de novo* variant class, the greater its overrepresentation among affected females.[5] This suggests that, on average, females accumulate more risk than males before being ascertained as ASD cases.

Male-female differences are less clear in the context of ASD's common, inherited genetic influences, which constitute the majority of genetic risk for ASD.[7] Given the findings above, we may expect elevated polygenic risk for ASD in female cases. That, however, has not been consistently observed.[1,8] This could be a function of statistical power, as the polygenic risk score (PRS) for ASD currently explains limited case-control variance on the liability scale (< 3%), and under 4,000 female cases are present in published ASD GWAS meta-analyses.[1,8]

Inconsistent observations of elevated polygenic risk for ASD in female cases could also reflect more complicated phenomena. One must make several assumptions in order to easily interpret a PRS comparison between male and female cases. First, one must assume equivalent genetic architecture between ASD as diagnosed in males (male ASD) and as diagnosed in females (female ASD). The differences in rare variant burden described above, along with preliminary evidence from studies of SNP heritability, already violate that assumption.[2–5,8] Second, one needs to assume that male ASD and female ASD have equivalent polygenic influences (a genetic correlation of 1). This is unclear at current sample sizes.[8] Even once that analysis becomes adequately powered, the correlation will be difficult to interpret. The male to female ratio in ASD increases with increasing case IQ, and this brings with it additional average differences in behavioral, cognitive, and medical comorbidities.[9] Any estimated genetic correlation between male and female ASD could accordingly conflate sex-based and phenotype-based heterogeneity.

In this study, we examined two alternate, complementary strategies for understanding the relationship between sex and inherited genetic risk for ASD. We first conducted a large sibling recurrence analysis, leveraging the Danish, nationally representative Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) resource. We then examined the relationship between sex and common polygenic risk for ASD in whole families, focusing on both affected and unaffected family members.

Results

FPE and Sibling Recurrence

Under the FPE model, one expects a greater aggregation of ASD risk in female cases than in male cases. In the context of inherited genetic risk, which is shared within families, that expectation extends to the family members of female cases. For example, we expect siblings of

female ASD cases to carry more risk for ASD than siblings of male ASD cases, regardless of whether they are categorically affected themselves.[10] Sibling recurrence is a particularly useful metric of inherited or familial risk. Full siblings share 50% of their segregating DNA variants and are typically close enough in age to share diagnostic environments. Shared diagnostic environment is particularly important when considering ASD recurrence. The estimated prevalence of ASD has increased over 30-fold over the last four decades[11], primarily due to diagnostic expansion.[9,12] Members of previous generations, particularly those able to live independently as adults, were far less likely to receive an ASD diagnosis in childhood than children born as of writing.[9,12] For this reason, inclusion of parents or aunts and uncles in familial recurrence analyses can complicate data interpretation. Our analysis was accordingly limited to siblings.

To further improve data interpretability, we stratified ASD cases based on presence or absence of co-diagnosed intellectual disability (ID). Despite sharing the majority of their rare variant influences[4], ID and ASD do not appear to share their common polygenic influences: as currently estimated, the genetic correlation between ID and ASD is not significantly different from zero.[13] Further, evidence suggests reduced SNP heritability for forms of ASD in which co-diagnosed ID is more common.[8] As (1) lower heritability predicts lower familial recurrence and (2) ascertained female ASD cases are more likely to have co-diagnosed ID, failing to stratify by ID could render a male-female comparison difficult to interpret. Our recurrence analyses focused on ASD without co-diagnosed ID (from here: *ASDnoID*), and used ID without co-diagnosed ASD (from here: *IDnoASD*) as a comparison group. We also included individuals with ID alone as a negative control. We excluded individuals with diagnoses of both ASD and ID (approximately 15% of ASD cases in Denmark), as there were too few cases in that group for an independent sibling recurrence analysis. It should be noted that the true ID rate in cases is likely much higher. If consistent with the rate of ID in ASD cases in the United States or the United Kingdom, it would be approximately 40% over this diagnostic period.[12] Intellectual disability among people with ASD is typically underreported in medical record and registry data, as notation of co-morbidities is an area of inconsistent clinical practice.

The Danish Psychiatric Central Research Register and the Danish National Patient Register are unique resources, well suited to careful consideration of sibling recurrence. They are complete until 2012 and 2013, respectively, and contain medical record data on the entire Danish population born between May 1, 1981 and December 31, 2005 (n = 1,472,762). We linked the psychiatric and patient registers to find all Danish families with two or more full siblings born during this time period. We identified 94,790 such families. We then identified the families with at least one child with *ASDnoID* or *IDnoASD*. This analysis included all diagnosed *ASDnoID* and *IDnoASD* cases in this population during this period. When a family included more than one affected child, we selected one at random to be the 'index case' (from here: cases). We analyzed one sibling per family; if the family included more than one sibling, we selected one at random for inclusion in the analysis. We examined ASD and ID diagnoses in the selected siblings. As the focus of the analysis was recurrence of ASD and ID, and any selection among siblings was performed at random, sibling selection was not diagnosis dependent (i.e., if the

family included a sibling with ASD and a sibling without, either could be selected, and with equal probability). A detailed description of this process can be found in the Online Methods.

We investigated whether siblings of female cases of *ASDnoID* (*n* = 1,707 siblings) have higher risk for ASD and/or ID themselves than the siblings of male cases of *ASDnoID* (*n* = 6,270 siblings). By requiring cases to have only one diagnosis, we were adequately powered to examine co-occurring ASD and ID (*ASDandID*) as an outcome in the siblings. In siblings, there were accordingly three potential outcomes: *ASDnoID*, *ASDandID*, and *IDnoASD*. We estimated sibling risk by comparing diagnosis rates in the siblings to diagnosis rates in age and sex matched controls, drawn at random from the Danish population. To increase power, we used 2:1 control to case matching. We followed the same procedures for siblings of female cases of *IDnoASD* (*n* = 506 siblings) and siblings of male cases of *IDnoASD* (*n* = 811 siblings).

The primary results are presented in Fig. 1. An odds ratio (OR) of more than one suggests that case siblings were more likely to receive a diagnosis than age and sex matched individuals from the general population. For example, siblings of female *ASDnoID* cases were approximately seven times as likely (OR = 7.19, 95% CI = 5.09-10.09) to receive a diagnosis of *ASDnoID* themselves than a general population individual. For siblings of male *ASDnoID* cases, there was a nearly four-fold (OR = 3.76, 95% CI = 3.10-4.54) increase in risk. In fact, while all siblings of *ASDnoID* cases were at increased ASD risk ($P < 1.34 \times 10^{-4}$ for all comparisons), the siblings of female *ASDnoID* cases were at even greater risk than the siblings of male *ASDnoID* cases ($P < 0.01$ for both comparisons). This is consistent with expectations of the FPE. We only compared risk between siblings of female and male cases if both sibling groups showed elevated risk against the general population. This is akin to only testing for an interaction in the presence of significant main effects.

The pattern was strikingly different for the siblings of *IDnoASD* cases. First, neither siblings of female cases (*n* = 506, *ASDandID*: OR = 2.00, 95% CI = 0.12-32.07, *ASDnoID*: OR = 2.01, 95% CI = 0.80-5.12) nor siblings of male cases (*n* = 811, *ASDandID*: OR = 6.02, 95% CI = 0.63-57.95, *ASDnoID*: OR = 1.49, 95% CI = 0.79-2.80) showed increased risk for ASD (with or without co-diagnosed ID) at these sample sizes. As increased risk for ASD could not be detected, we did not test for a difference in ASD risk between siblings of female versus male *IDnoASD* cases. The siblings of *IDnoASD* cases were, however, at significantly increased risk for *IDnoASD* themselves ($P < 3.13 \times 10^{-6}$ for both comparisons). This was true for both siblings of male cases and the siblings of female cases. Sibling risk of *IDnoASD* recurrence did not differ by the sex of the *IDnoASD* case ($P = 0.12$).

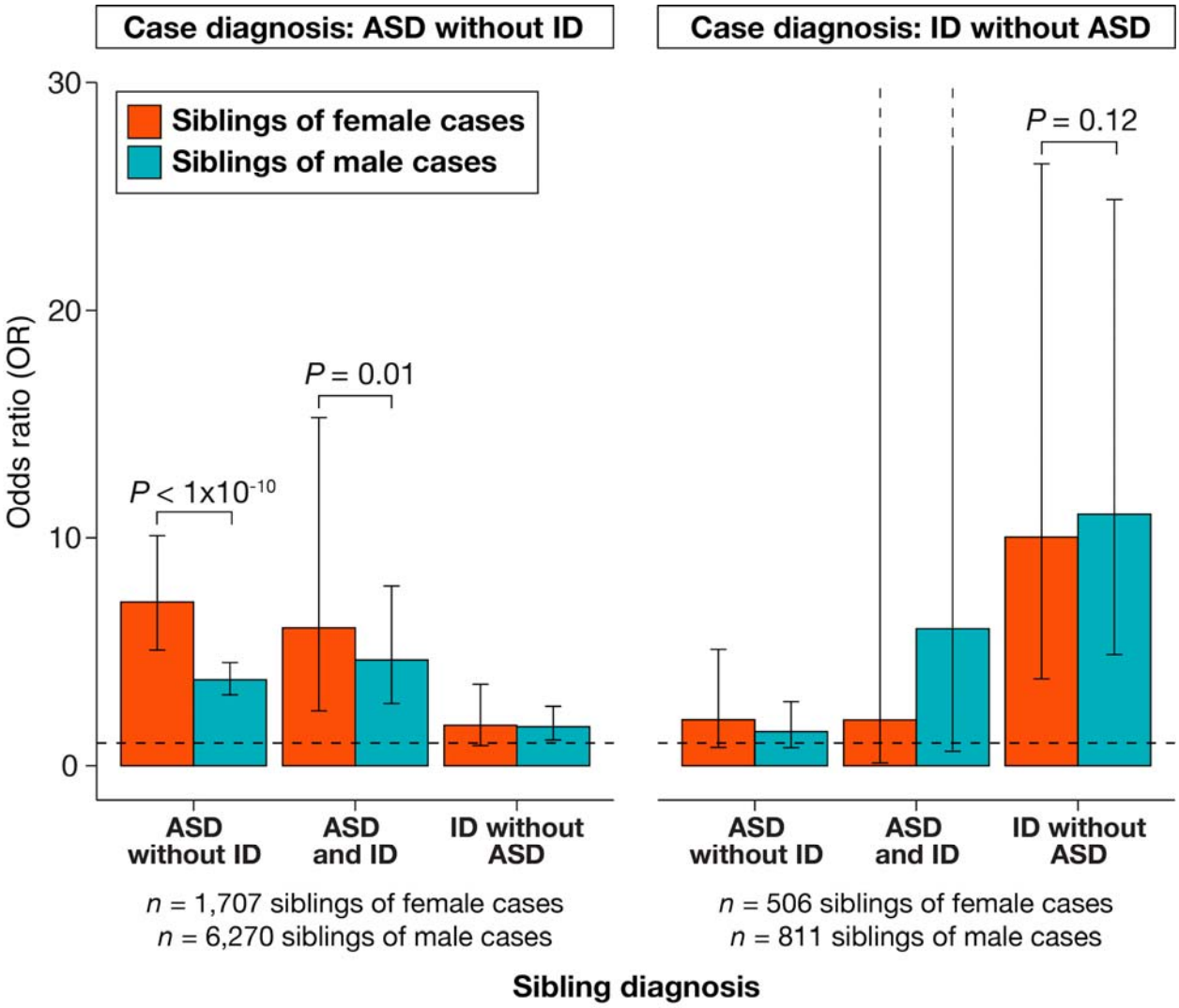

Figure 1. Sibling recurrence of ASD and ID. Red bars represent odds ratios (OR) for siblings of female cases and teal bars represent ORs for siblings of male cases. ORs indicate the increase in risk for each diagnosis among siblings of cases, as compared to age and sex matched controls. Error bars represent 95% confidence intervals. *P*-values are from a Wald test to determine whether ORs are significantly different from one another. *P*-values for the male-female comparison were only calculated when both ORs were significantly different from 1.

We were not statistically powered to simultaneously consider sex of the case and sex of the sibling. However, in an analysis of risk to male versus female siblings of all ASD cases, risk did not differ meaningfully by sex of the sibling when using a sex-specific general population rate (Supplementary Fig. 1).

FPE and ASD parents

We next examined the FPE in two genetically characterized ASD cohorts: the Simons Simplex Collection (SSC) and the Simons Foundation Powering Autism Research for Knowledge (SPARK) cohort. The SSC consists of families with one affected child and two confirmed,

unaffected parents.[14] SPARK includes families with a variety of structures (see Online Methods).

Parent-child designs present an opportunity to examine the role of the FPE in parents of cases, as well as in ASD cases themselves. We expect parents of ASD cases to have greater than average risk for ASD, simply because they have a child with ASD. The parents, however, are usually categorically unaffected. Some ASD studies, like the SSC, screened parents for ASD and ASD-like symptomatology. If a parent met criteria for an ASD diagnosis, or had an obvious and substantial concentration of ASD-like traits, the family could not participate in the study.[14] Families with ASD-diagnosed parents can participate in SPARK, but we excluded these families from our analysis (Online Methods). SPARK parents remaining in the analysis could still have a substantial aggregation of ASD symptomatology. However, as the parents have found a partner, had children, and registered for a research study, there are limits to the functional impairment that might come with those symptoms.

We expect mothers and fathers of children with ASD to carry elevated ASD risk relative to the general population. To estimate this increased risk, we integrated the SSC and SPARK data with a large general population cohort, the UK Biobank (UKB).[15] Using standard deviations (SD) on the UKB ASD PRS distribution as our scale, we then estimated the burden of common polygenic risk for ASD in all European ancestry parents in SPARK and SSC, as well as in ancestry matched controls from UKB. As expected, parents of ASD cases carried more genetic risk for ASD than controls (0.23 SD, $P = 1.9 \times 10^{-75}$, Fig. 2).

Under a FPE model, mothers would on average be able to carry more ASD risk than fathers before meeting ASD case criteria. Consistent with FPE expectations, we found that mothers of ASD cases carried significantly more polygenic risk for ASD than fathers of ASD cases (n = 7,436 mothers; n = 5,926 fathers, 0.09 SD, $P = 7.0 \times 10^{-7}$, Fig. 2). The increase in ASD PRS in ASD mothers compared to females in the general population was about 50% greater than the increase in ASD PRS in ASD fathers compared to males in the general population. This mother-father difference is present independently in both SSC (n = 2,061 mothers, n = 2,079 fathers, $P = 8.0 \times 10^{-3}$) and SPARK (n = 5,375 mothers, n = 3,847 fathers, $P = 5.2 \times 10^{-5}$). It is also present when comparing full trios: families where both parents are present in the dataset (n = 4,809 complete trios, $P = 1.4 \times 10^{-5}$). Further, while ASD cases had significantly greater PRS for ASD than their unaffected mothers on average (n = 7,628, 0.09 SD, $P = 1.2 \times 10^{-8}$, Fig. 2), that elevation was strikingly similar to the elevation observed between mothers and fathers.

There is no sex difference in ASD PRS in UKB (P = 0.15), as expected for any general population analysis featuring an autosomally-constructed PRS. The mother-father difference in ASD PRS we observed arises as a function of ascertaining families with an ASD proband.

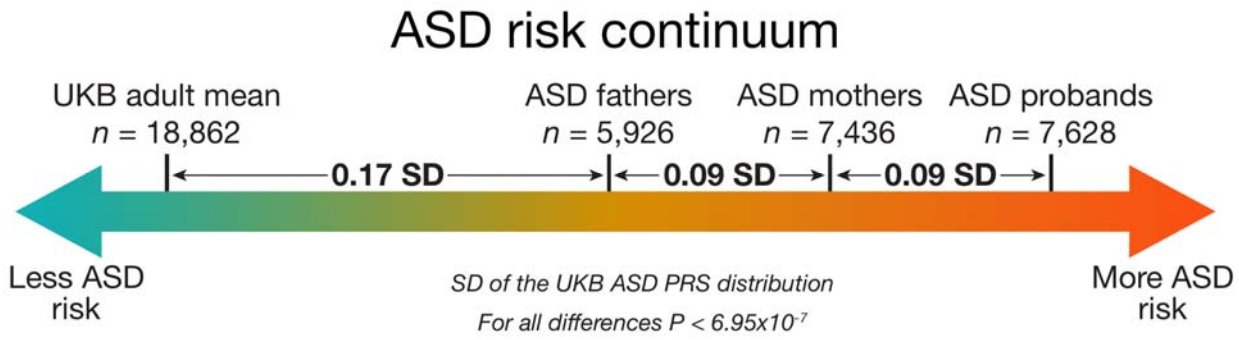

Figure 2: The continuum of ASD polygenic risk in the general population and families with an ASD case. Between group differences in polygenic score for ASD and *P*-values from linear regression comparing group polygenic scores while controlling for 15 principal components of ancestry. ASD groups are combined across the SSC and SPARK collections. Autosomal polygenic risk scores were calculated using weights from a GWAS of ASD cases (*n* = 19,870) and controls (*n* = 39,078) from the iPSYCH consortium in Denmark (Online Methods). Group differences are standardized using the UK Biobank ASD PRS distribution.

FPE and the polygenic transmission disequilibrium test (pTDT)

The polygenic transmission disequilibrium test (pTDT) compares polygenic risk between parents and their children. It leverages the expectation that, in a random sample of parent-child trios, the mean of the children's PRS for any trait will equal the mean of the mid-parent PRS (defined as the average of the mothers' and fathers' PRS). Ascertainment for a phenotypic deviation between children and parents, for example sampling children with ASD and parents without ASD, breaks that expectation, and allows one to identify polygenic risk factors that are associated with the ascertained outcome. We have previously shown that children with ASD, on average, substantially over-inherit their parents polygenic risk for ASD, as well as for schizophrenia and increased educational attainment.[1]

Larger ASD data sets, in conjunction with a new and better powered ASD PRS, allow us to revisit pTDT in light of the differential parental polygenic risk (Fig. 2). The difference in average ASD PRS between case mothers and case fathers changes our understanding of the mid-parent PRS. On average, male siblings of children with ASD are now expected to inherit more risk for ASD than is carried by their fathers (Fig. 3). To the extent that the mean difference in parental PRS reflects a sex difference in ASD risk tolerance, male siblings have substantially increased risk compared to female siblings. The difference in ASD PRS between ASD case mothers and fathers should be better tolerated in female siblings than in male siblings. The average mid-parent risk is less than the average risk carried by unaffected mothers of ASD cases, meaning females can tolerate higher risk than that expected in female siblings.

To investigate the FPE throughout families affected by ASD, we identified families in SSC and SPARK that include: (1) an affected child, (2) two unaffected parents, and (3) an unaffected sibling and performed pTDT on male and female unaffected siblings (*n* = 1,519 males, *n* = 1,611 females, Online Methods). We found that male unaffected siblings significantly under-inherit

their parents' polygenic risk for ASD (P = 0.03, Fig. 3). This is consistent with an average requirement for their PRS to decline from the mid-parental PRS to around that of their unaffected fathers, in order to remain unaffected themselves. We did not see a deviation from expectation in female siblings (P = 0.39, Fig. 3). While this is consistent with the FPE, the difference in transmission between male and female siblings is not statistically significant, and should be re-investigated with larger samples.

We last examined whether, using a new and better-powered ASD PRS, there is a difference in common polygenic burden between male and female cases. We did not detect any difference (P = 0.30). As noted in the introduction, interpretation of this null finding is complicated by several differences in ASD as it is ascertained in males and females, including: (1) a greater fraction of diagnosed females meeting criteria for ID; and (2) compared to males with ASD, diagnosed females have approximately twice the rate of high-impact *de novo* variants. Comparing the common variant architecture of male and female ASD accordingly involves many additional, unknown points of variation. To remove at least one source of sex-associated variation, we split male and female cases by presence/absence of a high-impact *de novo* variant.

We used exome sequence data from SSC and SPARK to identify the subset of ASD cases carrying a high-impact *de novo* variant, specifically predicted to disrupt the function of a constrained gene (12% of cases across both cohorts; see Online Methods). We hypothesized that high-impact *de novo* variants and the FPE create differences in the amount of liability space remaining to be filled by common polygenic variation. These differences may create the following ordering of polygenic over transmission (lowest to highest): (1) male cases with a high-impact *de novo* variant (n = 436), (2 and 3) either female cases with a high-impact *de novo* variant (n = 159) or male cases without a high-impact *de novo* variant (n = 3,468), (4) female cases without a high-impact *de novo* variant (n = 757).

The pTDT results reflected this expected gradient (Fig. 3). Male probands with high-impact *de novo* variants had the lowest polygenic over-inheritance (0.08 SD, P = 0.10), which was not significantly different from mid-parent expectation and was similar to that of their unaffected mothers (0.06 SD from the mid-parent value). Female cases without a high-impact *de novo* variant had nearly three times the polygenic over-inheritance (0.23 SD, $P = 7.82 \times 10^{-11}$) of male cases with a high-impact *de novo* variant (P = 0.02).

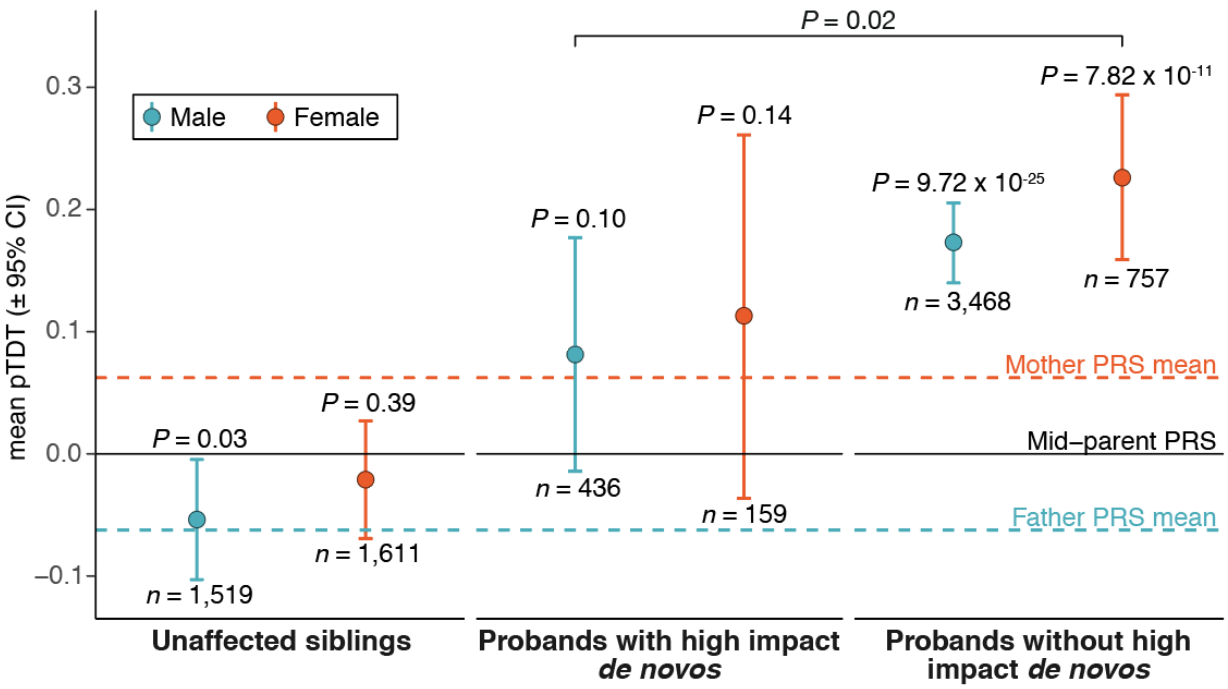

Figure 3: Polygenic transmission disequilibrium in ASD cases and unaffected siblings. Transmission disequilibrium standardized to the mid-parent PRS distribution with error bars denoting 95% confidence intervals. *P*-values are from a two-sided, one-sample *t*-test and estimate the probability that polygenic deviation is equal to 0. Cases and controls are combined across SSC and SPARK cohorts. The mother and father PRS mean lines are the mean values from pTDT of each parent against the mid-parent expectation (symmetric by definition). Summary statistics for the PRS are from a GWAS of ASD cases (*n* = 19,870) and controls (*n* = 39,078) from the iPSYCH consortium in Denmark (Online Methods).

Discussion

These results highlight the complicated but consistent relationship between sex and genetic risk for ASD. Evidence from multiple types of genetic risk, and multiple members of families affected by ASD, supports a female protective effect model, in which females have a higher liability threshold for receiving a diagnosis of ASD. We note that, in this analysis, female protection and male risk are one and the same. With only two categories and no insight into mechanism they are in fact indistinguishable.

We also note that polygenic risk for ASD is, in the general population, associated with many positive traits.[1,8,17] Dozens of studies have noted a positive, general population correlation between polygenic risk for ASD and greater educational attainment, stronger reasoning ability, and many other beneficial attributes in a cognitively-demanding economy. In females, the ability to tolerate more ASD risk without manifesting the most isolating elements of diagnosed ASD can benefit individuals, families, and communities.

While one may be tempted to quantify a formal expectation of ASD's genetic architecture under specified circumstances (e.g. female with a high-impact *de novo* variant; male without), such expectations would depend on a stable, or at least fairly predictable, phenotype. ASD, as

currently diagnosed, is neither. There are predictable elements of sex by phenotype interaction in diagnosed cases, for example escalating male to female ratio with increasing case IQ.[9] However, even after conditioning on IQ, one is left with residual phenotypic associations to sex among ascertained cases. For example, females are on average diagnosed later than boys.[12] Similarly, sex differences in genetic architecture remain after conditioning on presence/absence of a strong acting *de novo* variant. Across individuals with ASD, *de novo* variant count is associated with variant impact: as *de novo* variant count increases, so does their average effect size contribution to ASD.[1] Fewer of the variants are benign; more are likely clinically returnable.

We do not know what renders females more tolerant of ASD's genetic risk factors. Further, we do not know what, if anything, the mechanisms underlying that tolerance have in common with ASD genetic risk. Analysis at the molecular level will be necessary to address that question. At the statistical level, assuming adequate phenotypic stability and characterization, increasing sample sizes will lead to increasingly clear male-female differences. Future studies can further explore this axis of heterogeneity in ASD.

Online Methods

Identifying families in Danish Registry Data

The Danish Psychiatric Central Research Register and the Danish National Patient register, complete until 2012 and 2013, respectively, contain medical record data on the entire Danish population born between May 1, 1981 and December 31, 2005 (*n* = 1,472,76). The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH) consortium has established a large Danish population-based psychiatric Case–Cohort sample (iPSYCH2012) from this data to investigate the genetic and environmental architecture of severe mental disorders[18] . In this work, we focus specifically on ASD cases (*n* = 16,146), defined as individuals with ICD-10 codes F84.0, F84.1, F84.5, F84.8 or F48.9, as well as ID cases (*n* = 4,727), defined as individuals with any ICD-10 codes from F70-F79. Controls were population representative, randomly sampled individuals from the Danish population (*n* = 30,000). Controls may have psychiatric disorders, with prevalence levels amongst controls matching those seen in the Danish general population.

The iPSYCH2012 cohort contains medical diagnoses, prescribed medicine, and social and socioeconomic data for 449,882 individuals, and their first-degree relatives. Of those, 39,491 individuals had a missing identification number for one or both of their parents or were missing phenotypic sex. In total, there were 410,391 individuals with first degree relatives for which we had phenotypic sex, and an identification number for both parents. Amongst these 410,391 individuals, we identified 274,837 families. We further subset these families to those with more than one offspring (*n* = 94,790 families).

Comparing risk for NDDs between siblings of NDD cases and controls, by NDD case sex

For each family, we selected an index case based on two criteria: 1) sex (male or female), and 2) neurodevelopmental diagnosis (*ASDnoID*, *ASDandID*, or *IDnoASD*). Families without an index case were not considered. If more than one child in a family met the given criteria, one was randomly selected as the index case, with each offspring having an equal probability of being selected as the index case.

We then selected one sibling per index case. If an index case had more than one sibling, one was randomly selected, with each sibling having an equal probability of being selected. Selected siblings were subset to those born between 1981 and 2005. Each of these siblings were matched with two age-and sex-matched Danish population representative controls. All siblings of index cases were removed from the control cohort before being matched.

We then ran logistic regressions, NDD case status ~ $1_{\text{SIB OF CASE}}$ (where $1_{\text{SIB OF CASE}}$ is an indicator variable for whether the individual was the sibling of an NDD case (= 1) , or an age and sex matched control (= 0)), to investigate whether siblings of index cases have an increased risk for *ASDnoID*, *ASDandID*, and *IDnoASD* compared to age and sex matched controls.

ORss for increased risk with sibling case status are the exponentiated effect size for the association between sibling case status and diagnosis of a psychiatric disorder. To compare the ORs between siblings of female and male cases, we conducted a Wald test. The Wald test determines whether ORs (from the above described logistic regressions) are significantly different from one another.

This analysis was run for six types of index case: (1) female *ASDnoID*, (2) male *ASDnoID*, (3) female *ASDandID*, (4) male *ASDandID*, (5) female *IDnoASD* and (6) male *IDnoASD*.

Comparing risk for NDDs between siblings of NDD cases and controls, by NDD case sex

We conducted a similar analysis to the above to compare sisters with brothers of ASD cases. Previously, we investigated sibling risk by index case sex. Here, we investigate sibling risk by sibling sex. For each family, we selected an index ASD case regardless of sex and comorbid ID status. For each index case, we randomly selected a sibling, each with equal probability of selection. We then split the selected siblings by sex, into sisters and brothers of ASD cases.

Selected siblings were subset to those born between 1981 and 2005. Each of these siblings were matched with two age and sex matched Danish population representative controls (*n* = 30,000). All siblings of index cases were removed from the control cohort before being matched.

We then ran logistic regression, NDD case status ~ $1_{\text{SIB OF CASE}}$ (where $1_{\text{SIB OF CASE}}$ is an indicator variable for whether the individual was the sibling of an NDD case (= 1) , or an age and sex

matched control (= 0)), for sisters and brothers separately to investigate whether they have an increased risk for *ASDnoD*, *ASDandID*, and *IDnoASD* compared to age and sex matched controls.

ORs for increased risk with sibling case status are the exponentiated effect sizes for the association between sibling case status and diagnosis of a psychiatric disorder. To compare ORs between sisters and brothers of ASD cases, we conducted a Wald test.

Danish genotype data imputation

The iPSYCH2015 sample is an extension of the iPSYCH2012 sample expanding the birth cohorts by 3 years up to 2008 and extending the follow up to 2015, as well as drawing another 20,000 random samples for the random population subcohort. The new additional subsample is called iPSYCH2015i. Details of the sample, genotyping and call sets can be found in prior iPSYCH publications.[18,8,19]

Briefly, DNA was extracted from Guthrie cards in the Danish Neonatal Screening Biobank at Staten Serum Institute (SSI) and whole genome amplified. The two subsamples, iPSYCH2012 and iPSYCH2015i, were processed independently. Genotyping of the iPSYCH2012 sample was performed in 26 waves at the Broad Institute of Harvard and MITusing the PsychChip array from Illumina and the iPSYCH2015i sample was genotyped on the Global Screening Array v2 at the SSI.

Two stages of pre-imputation QC were conducted. In the first stage, we performed a near default Ricopili QC.[20] First SNPs with a call rate < 0.95 were removed. Then sample QC was run keeping individuals with a call rate in cases or controls ≥ 0.95 and an autosomal heterozygosity deviation FHET within +/- 0.20 of cases or controls. Subsequently, we ran marker QC. We retained markers with call rate ≥ 0.98, difference in missingness ≤ 0.02 between cases and controls, with minor allele frequency (MAF) ≥ 0.01 , Hardy-Weinberg equilibrium (HWE) in controls (P-value ≥ 1.0×10^{-6}) and HWE in cases (P-value ≥ 1.0×10^{-10}). See https://sites.google.com/a/broadinstitute.org/ricopili/preimputation-qc for further details.

The second stage of pre-imputation QC was targeted at batch effects. In iPSYCH2012 we considered three types of potential batch effects: pre-processing plate, array plate and wave, and in iPSYCH2015i we considered pre-processing plate, array plate, and array batch. We evaluated batch effects using unrelated, ancestry matched individuals in order to avoid confounding batch effects with population stratification or cryptic relatedness. For each of the three batch types, we looped over batches, performing a GWAS of each batch against the remaining batches. Association testing was conducted using PLINK v1.90b4. The exclusion of SNPs strongly associated with any of the batch types was based on the minimum P-value across all associations per batch type. The P-value cut-off for the wave and array batch was minimum $P < 2.0 \times 10^{-10}$, and for pre-processing plate and array plate, minimum $P < 2.0 \times 10^{-12}$.

Imputation was performed separately for the two samples following Ricopili defaults prephasing using Eagle v2.3.5[21] and imputation using Minimac3.[22] As reference we used the public part of the Haplotype Reference Consortium[23] (EGAD00001002729) prepared for the pipeline by the Ricopili team.[20]

Danish ASD GWAS

Our GWAS cases (*n* = 19,870) and controls (*n* = 39,078), are composed of iPSYCH2015 individuals with ASD and without ASD, respectively.

We defined sample ancestry based on a principal component analysis (PCA) using smartPCA[24,25]. We removed regions of extended linkage disequilibrium[26] (including the HLA region), and thinned the SNPs using PLINK2[26,27] by pruning those with pairwise $r^2 > 0.075$ in a window of 1000 SNPs with and step size of 100 SNPs, leaving roughly 30k markers.

Using PLINK's identity by state analysis, we identified pairs of samples with > 0.2, and excluded one sample from each pair at random (with a preference for keeping cases). We restricted the cohort to individuals of European ancestry defined as being within an ellipsoid in the space of principal components (PCs) 1-3. The ellipsoid was centered and scaled using the mean and standard deviation of individuals having all parents and grandparents born in Denmark according to national registries. We conducted a second PCA on these individuals and used the PCs as covariates for the association analysis.

We conducted association analyses separately in iPSYCH2012 and iPSYCH2015i using PLINK on the imputed dosage data, and controlling for the first ten PCs. We meta-analyzed the results of the two ASD GWAS using METAL[28] (July 2010 version) with an inverse variance weighted fixed effect model.[29]

SSC Imputation

The imputation of SSC has been described previously[1]. Note that the SSC cohort only includes unaffected parents and a single ASD proband. A single unaffected sibling per family is included in analysis; if there are multiple in a family, the sibling closest in age to the proband (SSC: "designated sibling") is included.

SPARK Imputation

SPARK data were processed, restricted to individuals of European ancestry, and imputed using the picopili pipeline[30] (https://github.com/Nealelab/picopili), which is an adaptation and extension of Ricopili[20] for family data. Phasing and imputation were conducted using SHAPEIT[31] and IMPUTE2[32], respectively, using Haplotype Reference Consortium[23] (HRC) data and genome build hg19. Genotypes were called for 7,124,628 autosomal SNPs (minimum posterior probability > 0.8), with a genotyping rate of 0.995 across 16,965 samples of European ancestry.

We removed SPARK parents with an ASD diagnosis from analysis. We included all probands from multiplex families as well as all unaffected siblings.

De novo variant analysis

We downloaded gVCFs generated by GATK for 27,270 individuals from SFARIbase (/SPARK/Regeneron/SPARK_Freeze_20190912/Variants/GATK/). All gVCFs were generated with GATK v4.1.2.0 HaplotypeCaller using default thresholds and based on hg38 reference and target files provided by Regeneron (genome.hg38rg.fa & xgen_plus_spikein.b38.bed respectively). New quality scores, lenient processing of VCF files, and 100bp padding for intervals were also used. We then performed joint calling of these 27,270 sample gVCFs via GATK to produce one unified vcf for the SPARK cohort. Subsequent variant filtering QC and *de novo* variant detection were carried out using consistent thresholds with those described previously.[4]

We identified the ASD probands in SSC and SPARK who carried a *de novo* variant in a class previously associated with ASD risk. These variants constitute three groups: 1) protein-truncating variants to genes intolerant of heterozygous loss of function variation (constrained gene)[1], 2) copy number variants (deletions or duplications) affecting at least one constrained gene[1] and 3) predicted protein-altering missense variant in a constrained gene (missense class B variant[4]). Collectively, 11.6% of SSC probands carry at least one of these variants, while 12.2% of SPARK probands carry at least one. Across SSC and SPARK, 11.2% of male probands carry at least one of these variants, while 17.4% of female probands carry at least one.

Ancestry definition in SSC, SPARK and UKB

We randomly selected 20,000 samples from UKBB to serve as the population control cohort. Using PLINK (version 1.9), we then constructed a merged file with these genotyped controls, SSC (n = 10,206), SPARK (n = 16,965) and HapMap 3[33] (n = 988) for the purpose of defining ancestry. We retained SNPs with MAF > 0.01 and per SNP missingness < 0.25%. Of the remaining SNPs, we randomly sampled 10,000 to improve the computational efficiency of calculating PCs. We then used PLINK to calculate the PCs. To define ancestry, we projected all 48,159 samples into their joint principal component space and selected a sub-sample of our cases and controls that clustered with Europeans in HapMap (-0.002 < PC1 < 0.003, -0.004 < PC2 < 0.003).

We then calculated PCs in this European ancestry subset of UKB, SSC and SPARK. First, we retained SNPs with MAF > 0.01 and missingness < 1%. Then, we performed LD pruning using PLINK to retain SNPs in approximate linkage equilibrium (--indep-pairwise 50 5 0.15). Next, we removed SNPs in 24 regions of long-range LD (mean partition size: 5.5Mb[26]). We then used PLINK to perform PCA on the remaining 95,509 SNPs and used the first 15 PCs for downstream analyses to control for ancestry.

Generation of polygenic risk score

We used LDpred[34] (version 1.0.11) and the marginal effect sizes from the iPSYCH2015 ASD GWAS to generate a polygenic risk score, using the infinitesimal model, European ancestry subset of Hapmap 3 for LD reference, and an LD radius of 384 SNPs (per LDpred guidance). The weights from LDpred were used to calculate per sample ASD PRS using linear scoring in PLINK. There were 630,583 markers in common between the genotypes and the markers in the iPSYCH2015 ASD GWAS summary statistics, all of which were used in the polygenic risk score.

Polygenic risk comparisons

We performed two classes of analyses to compare polygenic burden between groups. The first is a within-family polygenic transmission disequilibrium test[1], where a t-statistic of the deviation of the offspring's polygenic risk from the mean parent expectation is compared to the null hypothesis of 0, using a one-sample *t*-test. This approach was performed for all comparisons in Figure 3. There is no restriction of ancestry in this analysis as comparisons are within family transmission tests. Polygenic deviations are scaled by the standard deviation of the distribution of mid-parent PRS for all families with a sequenced proband in SSC+SPARK. The comparison of pTDT values between groups in Figure 3 is performed as a 2-sample *t*-test of each pTDT deviation distribution. The second approach is a between-group comparison, where the PRS between two groups is compared using linear regression while controlling for PCs, specifically: ASD PRS ~ group indicator + PCs1-15. Here, only samples of European ancestry and their PCs are used (as discussed above in "Ancestry definition"). This approach was performed for comparisons in Figure 2. The between group differences in PRS are scaled by the standard deviation of the distribution of ASD PRS in the UK Biobank controls.

Acknowledgements

This work was supported by the Autism Science Foundation (ASP 001 to SJS, ASP 002 to SLB, ASP 003 to EBR) and the NIMH (RMH111813A to EBR; U01MH111662 to SJS). DJW was supported by NLM grant T15LM007092. The iPSYCH team was supported by grants from the Lundbeck Foundation (R102-A9118, R155-2014-1724, and R248-2017-2003), the EU H2020 Program (Grant No. 667302, "CoCA" to ADB), NIMH (1U01MH109514-01 to ADB) and the Universities and University Hospitals of Aarhus and Copenhagen. The Danish National Biobank resource was supported by the Novo Nordisk Foundation. High-performance computer capacity for handling and statistical analysis of iPSYCH data on the GenomeDK HPC facility was provided by the Center for Genomics and Personalized Medicine and the Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark (grant to ADB). This research has been conducted using data from UK Biobank, a major biomedical database, under Project 31063. The authors would like to deeply thank all participants in the cohorts included in this analysis.

Author contributions

EMW, DJW, JG, AR, JMF, WT, CEC, NB, CvdM, RKW, FKS, DSP, JBG conducted data analysis. EMW, DJW, and EBR wrote the paper. DMH, PBM, MJD, MET, ADB, and EBR supervised data analysis. EMW, DJW, SJS, SLB, and EBR designed the study.

Declarations

The details of the IRB/oversight body that provided approval or exemption for the research described are: This study was reviewed and approved by Partners Human Research of Partners HealthCare. The study name is Molecular Study of Cognitive and Behavioral Variation (IRB: 2015P002376). The Principle Investigator is Elise Robinson.

Competing interests

DSP is an employee of Genomics plc. All the analyses reported in this paper were performed as part of DSP.'s previous employment at the Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA, and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. All other authors declare no competing interests.

Data Availability Statement

iPSYCH: Data availability is limited due to the sensitive nature of the data. The iPSYCH Consortium is working with GDPR compliant models for remote access; please contact authors Preben Bo Mortensen and Anders D. Børglum for more details.

The imputed SPARK dataset used in this analysis has been given to the Simons Foundation Autism Research Initiative (SFARI) for public distribution. Scientists wishing to access the data set can do so through application to SFARI.

HapMap 3 data are available here:
https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html

UK Biobank: approved researchers can access the data by applying at
https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access

Haplotype Reference Consortium. Researchers can impute genotypes using the HRC reference data at the following sites: https://imputation.sanger.ac.uk/,
https://imputationserver.sph.umich.edu/

References

1. 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).
2. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
3. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
4. 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).
5. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).
6. Satterstrom, F. K. *et al.* Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
7. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
8. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
9. Fombonne, E. Epidemiological surveys of autism and other pervasive developmental disorders: an update. *J. Autism Dev. Disord.* **33**, 365–382 (2003).
10. Robinson, E. B., Lichtenstein, P., Anckarsäter, H., Happé, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl. Acad. Sci. U.*

S. A. **110**, 5258–5262 (2013).

11. Boat, T. F. *et al.* Prevalence of Autism Spectrum Disorder. in *Mental Disorders and Disabilities Among Low-Income Children* (National Academies Press (US), 2015).
12. Maenner, M. J. *et al.* Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. *MMWR Surveill. Summ.* **69**, 1–12 (2020).
13. Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
14. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
15. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
16. Messinger, D. *et al.* Beyond autism: a baby siblings research consortium study of high-risk children at three years of age. *J. Am. Acad. Child Adolesc. Psychiatry* **52**, 300–308.e1 (2013).
17. 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).
18. Pedersen, C. B. *et al.* The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
19. Bybjerg-Grauholm, J. *et al.* The iPSYCH2015 Case-Cohort sample: updated directions for unravelling genetic and environmental architectures of severe mental disorders. *medRxiv* 2020.11.30.20237768 (2020).
20. Lam, M. *et al.* RICOPILI: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics* **36**, 930–933 (2019).

21. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
22. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
23. Huang, J. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
24. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
25. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
26. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* vol. 83 132–5; author reply 135–9 (2008).
27. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
28. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
29. Begum, F., Ghosh, D., Tseng, G. C. & Feingold, E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* **40**, 3777–3784 (2012).
30. Walters, R. K. *et al.* Trans-ancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat. Neurosci.* **21**, 1656 (2018).
31. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).
32. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, e1000529 (2009).

33. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

34. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).