nature genetics

Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population

Elise B Robinson[1–3,23], Beate St Pourcain[4,5,23], Verneri Anttila[1–3], Jack A Kosmicki[1–3,6], Brendan Bulik-Sullivan[1–3], Jakob Grove[7–10], Julian Maller[1–3], Kaitlin E Samocha[1–3,11], Stephan J Sanders[12], Stephan Ripke[1–3,13], Joanna Martin[1–3], Mads V Hollegaard[14], Thomas Werge[8,15,16], David M Hougaard[14], iPSYCH-SSI-Broad Autism Group[17], Benjamin M Neale[1–3,18], David M Evans[4,19], David Skuse[20], Preben Bo Mortensen[7,8,21], Anders D Børglum[7–9], Angelica Ronald[22], George Davey Smith[4] & Mark J Daly[1–3]

Almost all genetic risk factors for autism spectrum disorders (ASDs) can be found in the general population, but the effects of this risk are unclear in people not ascertained for neuropsychiatric symptoms. Using several large ASD consortium and population-based resources (total $n > 38{,}000$), we find genome-wide genetic links between ASDs and typical variation in social behavior and adaptive functioning. This finding is evidenced through both LD score correlation and *de novo* variant analysis, indicating that multiple types of genetic risk for ASDs influence a continuum of behavioral and developmental traits, the severe tail of which can result in diagnosis with an ASD or other neuropsychiatric disorder. A continuum model should inform the design and interpretation of studies of neuropsychiatric disease biology.

ASDs are a group of neuropsychiatric conditions defined through deficits in social communication, as well as restricted and repetitive interests. Consistent with traditional approaches to psychiatric phenotypes, most genetic studies of ASDs compare cases to controls to identify risk-associated variation. This approach has been highly productive—recent studies have linked common polygenic as well as *de novo* and inherited rare variation to ASD risk[1,2]. Common genotyped SNPs are estimated to account for at least 20% of ASD liability[1,3,4]. Contributing *de novo* variants are found in 10–20% of cases, but *de novo* mutations collectively explain less than 5% of overall ASD liability[1,5,6].

Almost all genetic risk factors for ASDs can be found in unaffected individuals. For example, most people who carry a 16p11.2 deletion, the most common large mutational risk factor for ASDs, do not meet the criteria for an ASD diagnosis[7]. Across healthy populations, there is also substantial variability in capacity for social interaction and social communication[8]. Although such phenotypic variation is well established, the genetic relationship between neuropsychiatric disorders and typical social and behavioral variation remains unclear. From the first published descriptions of ASDs, clinical and epidemiological reports have commonly noted subthreshold traits of autism in the family members of many diagnosed individuals[9,10]. Twin and family studies have suggested that these similarities are at least in part inherited and also suggest that traits and diagnosis are correlated genetically[11–13], but the correlation has yet to be estimated using measured genetic data.

We examined the association between genetic risk for ASDs and social and behavioral variation in the general population as well as the model through which genetic risk for ASDs is conferred. Traditional categorical psychiatric diagnoses (for example, yes/no for ASD) ignore the possibility of intermediate outcomes, long known to be relevant to phenotypes such as intellectual disability and IQ that are more easily quantified. Several studies have now associated copy number variants (CNVs) that create risk for neuropsychiatric disease with cognitive or educational differences in the general population[14,15]. *De novo* deletions at 16p11.2 were recently reported to confer a

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [4]Medical Research Council Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [5]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. [6]Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [7]Department of Biomedicine–Human Genetics, Aarhus University, Aarhus, Denmark. [8]Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Copenhagen, Denmark. [9]Centre for Integrative Sequencing, iSEQ, Aarhus University, Aarhus, Denmark. [10]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [11]Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [12]Department of Psychiatry, University of California, San Francisco, San Francisco, California, USA. [13]Department of Psychiatry and Psychotherapy, Charité, Campus Mitte, Berlin, Germany. [14]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. [15]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Mental Health Services Copenhagen, Copenhagen, Denmark. [16]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [17]A list of members and affiliations appears at the end of the paper. [18]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA. [19]Diamantina Institute, Translational Research Institute, University of Queensland, Brisbane, Queensland, Australia. [20]Behavioural Sciences Unit, Institute of Child Health, University College London, London, UK. [21]National Centre for Register-based Research, University of Aarhus, Aarhus, Denmark. [22]Department of Psychological Sciences, Centre for Brain and Cognitive Development, Birkbeck University of London, London, UK. [23]These authors contributed equally to this work. Correspondence should be addressed to E.B.R. (erobinson@atgu.mgh.harvard.edu) or M.J.D. (mjdaly@atgu.mgh.harvard.edu).

Received 3 September 2015; accepted 22 February 2016; published online 21 March 2016; doi:10.1038/ng.3529

quantitative effect on intelligence (resulting in an average reduction of 2 s.d. from the mean IQ of the parents), rather than creating risk for categorical (yes/no) intellectual disability[16]. The extent to which such patterns extend to social and behavioral traits is unknown and could substantially influence (i) the design and interpretation of biological studies of ASDs and other severe mental illnesses and (ii) the designation of therapeutic treatment thresholds. We aimed to resolve this question using multiple categories of genetic variation that create risk for ASDs, including common, inherited alleles as well as rare, *de novo* variants.

As in nearly all common diseases, common variant risk for ASDs is distributed across the genome, with many thousands of contributing loci of small effect[1,3]. The cumulative contribution of common SNPs to ASD risk (SNP heritability[17]) has been estimated using several methods, most recently and efficiently through linkage disequilibrium (LD) score regression. LD score regression makes use of genome-wide association study (GWAS) summary statistics to estimate SNP heritability[4]. The method can also be used to estimate the correlation between common variant influences on two phenotypes (genetic correlation, or r_g)[18]. As LD score correlation requires only GWAS summary statistics, genetic correlations can be estimated between distinct data sets and cohorts.

We used three data sets to examine the common variant association between ASDs and social and communication difficulties in the general population (**Supplementary Table 1**). First, traits of social and communication impairment were measured using the Social and Communication Disorders Checklist (SCDC) in the Avon Longitudinal Study of Parents and Children (ALSPAC), a general population cohort born from 1991–1992 in Bristol, UK[19,20]. The SCDC is a parent-rated quantitative measure of social communication impairment that is continuously distributed and has been extensively studied[21–23]. There is substantial trait overlap between the SCDC and canonical ASD symptomology (for example, "not aware of other people's feelings"), and children with ASDs on average have very high scores (indicating many difficulties) on the SCDC[21]. The measure does not include items on restricted and repetitive interests. For the purposes of this project, we used summary statistics from a published GWAS of the SCDC, administered when the children were 8 years old (n = 5,628)[23]. The SNP heritability of the SCDC was 0.17 (standard error (s.e.) = 0.09) using LD score regression, similar to the estimate derived from residual maximum-likelihood analysis using the software package GCTA (h^2_g = 0.24, s.e. = 0.07; n = 5,204)[23].

We correlated the genetic influences on the SCDC with those on diagnosed ASDs using ASD data from two large consortium efforts. The Psychiatric Genomics Consortium autism group (PGC-ASD) has completed a GWAS of 5,305 ASD cases and 5,305 pseudocontrols constructed from untransmitted parental chromosomes (Online Methods). Summary statistics from this GWAS are publicly available through the PGC website (see URLs). As a replication set, we recently completed an independent ASD case-control GWAS with 7,783 ASD cases and 11,359 controls from the Danish iPSYCH project (iPSYCH-ASD; Online Methods). Using LD score regression, we estimated that the liability-scale SNP heritability for PGC-ASD was 0.23 (s.e. = 0.03; assumed population prevalence of 1%), suggesting that approximately one-quarter of ASD liability reflects common genotyped variation. The estimated liability-scale SNP heritability for iPSYCH-ASD was 0.14 (s.e. = 0.03; assumed population prevalence of 1%). The genetic correlation between PGC-ASD and iPSYCH-ASD was 0.74 (s.e. = 0.07; $P < 1 \times 10^{-20}$), indicating similar common, polygenic architectures for ASDs diagnosed primarily in the United States and Denmark.

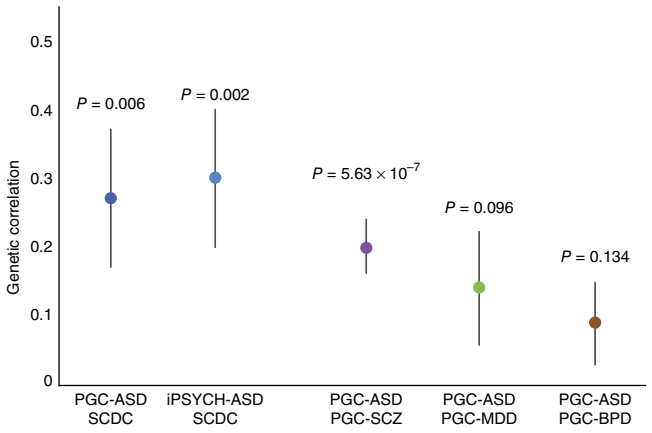

Figure 1 The genetic correlation between ASDs and pediatric social and communication difficulties in the general population. Genetic correlations are shown ±1 s.e.; *P* values indicate the probability that the true genetic correlation is 0. Genetic correlations were estimated using constrained-intercept LD score correlation. The individual correlations between PGC-ASD and PGC-SCZ (schizophrenia), PGC-MDD (major depressive disorder) and PGC-BPD (bipolar disorder) were modified from Bulik-Sullivan *et al.*[18].

The estimated genetic correlations between the SCDC in ALSPAC and the two ASD GWAS data sets are shown in **Figure 1**. The estimated genetic correlation between PGC-ASD and the SCDC was 0.27 (s.e. = 0.13; P = 0.006), suggesting that approximately one-quarter of the genetic influences on ASDs also influence the SCDC. The estimated genetic correlation between iPSYCH-ASD and the SCDC was similar (r_g = 0.30, s.e. = 0.10; P = 0.002), evidencing substantial and replicable etiological overlap between ASDs and typical variation in social and communication ability in childhood. The estimated genetic correlations between ASDs and the SCDC met or exceeded those previously estimated between PGC-ASD and each of PGC schizophrenia, PGC bipolar disorder and PGC major depressive disorder (**Fig. 1**). This suggests that ASDs are at least as strongly associated with variation in social and communication traits in the population as they are with several other categorically diagnosed psychiatric disorders. The observed genetic correlation between ASDs and the SCDC is similar to that estimated between type 2 diabetes and obesity, as well as other phenotypes that are strongly associated epidemiologically[18]. Many behavioral and cognitive features are captured in an ASD diagnosis: social communication impairment, restricted and repetitive interests, functional impairments and often co-occurring phenotypes that increase the probability of diagnosis (for example, intellectual disability or hyperactivity). The SCDC captures only traits of social and communication impairment, and the genetic association between ASDs and ASD traits might be stronger if more dimensions of the ASD phenotype were included among the traits measured. Supplemental analyses suggest that the estimated ASD-SCDC correlations are not driven by a negative association between ASDs and IQ (**Supplementary Table 2**).

We next aimed to examine whether *de novo* variant data similarly give evidence of a genetic relationship between ASDs and a continuum of behavioral outcomes in the population. The Simons Simplex Collection (SSC) is a sample of over 2,800 individuals with ASDs and their nuclear family members, with extensive data on unaffected siblings[24]. To our knowledge, the SSC unaffected siblings are currently the only deeply phenotyped control sample with *de novo* variant information available. The Vineland Adaptive Behavior Scales (Vineland) were used for the sequenced SSC cases (n = 2,497) and sibling controls (n = 1,861). The

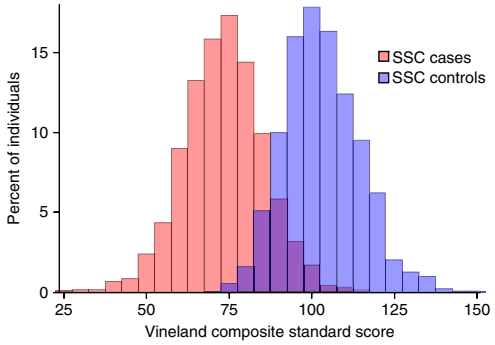

Figure 2 The distribution of Vineland score overlaps between SSC cases and controls. The Vineland composite standard score is normed, across ages, at a mean of 100 and an s.d. of 15 in the general population. The SSC case mean (73.3, s.d. = 12.2) is significantly lower than the SSC control mean (103.0, s.d. = 11.3; $P < 1 \times 10^{-20}$).

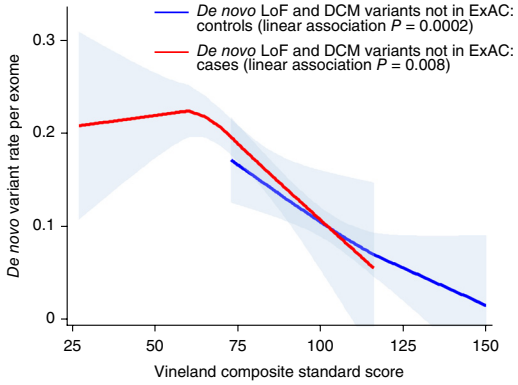

Figure 3 *De novo* variation influences a continuum of functional outcomes in ASD cases and controls. Natural (loess) association is shown for both cases and controls; *P* values are derived from Poisson regression. The shaded regions represent 95% confidence limits for association.

Vineland captures parent-rated variation in social, communication and daily living skills and normalizes these abilities to a mean of 100 and an s.d. of 15 in the population, corrected for age[25]. On average, individuals with ASDs have mean Vineland scores approximately 2 s.d. below the general population mean (SSC case mean = 73.3, s.d. = 12.17), consistent with the social and communication impairments definitional to diagnosis. The distribution of Vineland score is overlapping for cases and controls and is shown in **Figure 2**.

We examined the association between Vineland scores and *de novo* variant burden in individuals with and without ASDs. Our previous work demonstrated genotype-phenotype relationships using variant classes that are strongly associated with ASD risk[26]; subsequently, these analyses focused on (i) *de novo* loss-of-function (LoF) variants and (ii) *de novo* missense variants predicted to be damaging by the variant annotation program PolyPhen-2 and occurring in a gene known to be intolerant of heterozygous missense variation (DCM variants)[27,28]. We found that variants in one of these two categories could be identified in 22.1% of SSC cases and 13.2% of SSC unaffected siblings (LoF + DCM variant carrier ratio (CR) = 1.68; $P = 1.8 \times 10^{-11}$; Online Methods). To enhance signal, we further filtered the list of variants to remove *de novo* variants that were also seen in adult individuals in the Exome Aggregation Consortium (ExAC) resource (see URLs). The ExAC database includes 60,706 exomes. Variants absent from this reference panel, which is a proxy for standing variation in the human population, are more likely to be deleterious. For example, 18.3% of the *de novo* LoF + DCM variants in the SSC are found in the ExAC database, and, once these are removed, the relative *de novo* LoF + DCM variant burden in cases increases (CR for LoF + DCM variants not in ExAC = 1.91; $P = 7.6 \times 10^{-15}$).

The natural association between (i) *de novo* LoF + DCM variants not seen in the ExAC database and (ii) Vineland score in SSC cases and controls is shown in **Figure 3**. The LoF + DCM variant rate is predicted linearly by functional impairment in both cases ($P = 0.008$) and controls ($P = 0.0002$) using Poisson regression, controlling for sex. Cases and controls with equivalent quantitative levels of functional impairment, a key component of all psychiatric diagnoses, are highly similar with regard to *de novo* variant burden, suggesting that the current categorical clinical threshold is largely arbitrary with regard to the social and communication impairments captured by the Vineland. The strength of the association between LoF + DCM variant burden and case status ($P = 7.6 \times 10^{-15}$ without controlling for Vineland score) is only nominally significant after controlling for Vineland score ($P = 0.05$). The associations were weaker but similar without the ExAC filter (**Supplementary Fig. 1**).

These data strongly suggest that genetic influences on ASD risk—both inherited and *de novo*—influence typical variation in the population in social and communication ability. They also link clinically significant problems to impairments that are less likely to be ascertained. The results have major implications for genetic models of neuropsychiatric disorder risk. It is likely that inherited liability for ASDs is reflected in the behavioral traits of some family members of affected individuals. This links genetic and phenotypic burden in an intuitively consistent fashion with complex, continuously distributed polygenic disease risk. For traits such as height, it is simple to conceptualize a model in which tall parents (for example, those with a height 2 s.d. above the mean) are more likely to have a child who is very tall (for example, one with a height 3 s.d. above the mean). Historically, this concept has been more complicated in neuropsychiatric disorders. Despite extensive evidence, some have even questioned the role of inheritance given that the parents of individuals with ASDs or schizophrenia rarely carry a diagnosis themselves. These results suggest that familiality should be studied in a manner beyond a count of categorically affected family members and that trait variation in controls can provide insight into the underlying etiology of severe neurodevelopmental and psychiatric disorders. The behavioral influence of *de novo* and inherited genetic risk for ASDs can be quantified, and studies assessing continuous trait variation are likely better equipped to examine the phenotypic correlates of neuropsychiatric disease risk.

URLs. LDSC, http://www.github.com/bulik/ldsc; ALSPAC, http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary; Psychiatric Genomics Consortium (PGC), http://www.med.unc.edu/pgc/downloads; Exome Aggregation Consortium (ExAC), http://exac.broadinstitute.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Hyman, T. Lehner and N. Kanwisher for comments on the manuscript. We are extremely grateful to all the families who took part in this study, the

midwives for their help in recruiting them and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council, the Wellcome Trust (grant 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. Autism Speaks (7132) provided support for the analysis of autistic trait–related data in ALSPAC (to B.S.P.). This work was also supported by the Medical Research Council Integrative Epidemiology Unit (MC_UU_12013/1-9). This publication is the work of the authors, and E.B.R. and M.J.D. will serve as guarantors for the contents of this paper. The ALSPAC GWAS data were generated by the Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. E.B.R. was funded by National Institute of Mental Health grant 1K01MH099286-01A1 and Brain Behavior Research Foundation (NARSAD) Young Investigator grant 22379. We thank the families who took part in the Simons Simplex Collection study and the clinicians who collected data at each of the study sites. The iPSYCH project is funded by the Lundbeck Foundation and the universities and university hospitals of Aarhus and Copenhagen. Genotyping of iPSYCH and PGC samples was supported by grants from the Stanley Foundation, the Simons Foundation (SFARI 311789 to M.J.D.) and the National Institute of Mental Health (5U01MH094432-02 to M.J.D.). The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found on the ExAC website (see URLs). This work was also supported by a grant from the Simons Foundation (SFARI 307705; to S.J.S.).

AUTHOR CONTRIBUTIONS

E.B.R., B.S.P., V.A., J.A.K., B.B.-S., J.G., J. Maller, K.E.S., S.J.S., D.M.E., S.R., J. Martin, M.V.H., T.W., D.M.H., P.B.M. and A.D.B. generated data and/or conducted analyses. E.B.R., B.S.P., B.B.-S., B.M.N., J. Martin, D.S. and M.J.D. designed the experiment and tools. P.B.M., A.D.B., A.R., G.D.S. and M.J.D. supervised the research. E.B.R., B.S.P. and M.J.D. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Gaugler, T. *et al.* Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
2. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
3. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
4. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
5. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
6. Iossifov, I. *et al.* The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
7. Hanson, E. *et al.* The cognitive and behavioral phenotype of the 16p11.2 deletion in a clinically ascertained population. *Biol. Psychiatry* **77**, 785–793 (2015).
8. Plomin, R., Haworth, C.M. & Davis, O.S. Common disorders are quantitative traits. *Nat. Rev. Genet.* **10**, 872–878 (2009).
9. Kanner, L. Autistic disturbances of affective contact. *Nervous Child* **2**, 217–250 (1943).
10. Constantino, J.N., Zhang, Y., Frazier, T., Abbacchi, A.M. & Law, P. Sibling recurrence and the genetic epidemiology of autism. *Am. J. Psychiatry* **167**, 1349–1356 (2010).
11. Robinson, E.B. *et al.* Evidence that autistic traits show the same etiology in the general population and at the quantitative extremes (5%, 2.5%, and 1%). *Arch. Gen. Psychiatry* **68**, 1113–1121 (2011).
12. Lundström, S. *et al.* Autism spectrum disorders and autistic like traits: similar etiology in the extreme end and the normal variation. *Arch. Gen. Psychiatry* **69**, 46–52 (2012).
13. Ronald, A. *et al.* Genetic heterogeneity between the three components of the autism spectrum: a twin study. *J. Am. Acad. Child Adolesc. Psychiatry* **45**, 691–699 (2006).
14. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
15. Männik, K. *et al.* Copy number variations and cognitive phenotypes in unselected populations. *J. Am. Med. Assoc.* **313**, 2044–2054 (2015).
16. Moreno-De-Luca, A. *et al.* The role of parental cognitive, behavioral, and motor profiles in clinical variability in individuals with chromosome 16p11.2 deletions. *JAMA Psychiatry* **72**, 119–126 (2015).
17. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
18. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
19. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **42**, 111–127 (2013).
20. Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).
21. Skuse, D.H., Mandy, W.P. & Scourfield, J. Measuring autistic traits: heritability, reliability and validity of the Social and Communication Disorders Checklist. *Br. J. Psychiatry* **187**, 568–572 (2005).
22. Robinson, E.B. *et al.* Stability of autistic traits in the general population: further evidence for a continuum of impairment. *J. Am. Acad. Child Adolesc. Psychiatry* **50**, 376–384 (2011).
23. St Pourcain, B. *et al.* Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. *Mol. Autism* **5**, 18 (2014).
24. Fischbach, G.D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
25. Sparrow, S.S., Cicchetti, D.V. & Balla, D.A. *Vineland Adaptive Behavior Scales* (Pearson, 2005).
26. Robinson, E.B. *et al.* Autism spectrum disorder severity reflects the average contribution of *de novo* and familial influences. *Proc. Natl. Acad. Sci. USA* **111**, 15161–15165 (2014).
27. Samocha, K.E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
28. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

iPSYCH-SSI-Broad Autism Group collaborators:

Thomas D Als[7–9], Marie Baekvad-Hansen[14], Richard Belliveau[2], Ditte Demontis[7–9], Ashley Dumont[2], Jacqueline Goldstein[1–3], Jonas Grauholm[14], Christine S Hansen[14], Thomas F Hansen[8,15], Daniel Howrigan[1–3], Francesco Lescai[7–9], Manuel Mattheisen[7–9], Jennifer Moran[2], Ole Mors[8,24], Merete Nordentoft[8,25], Bent Norgaard-Pedersen[14], Timothy Poterba[1–3], Jesper Poulsen[14], Christine Stevens[2] & Raymond Walters[1–3]

[24]Research Department P, Aarhus University Hospital, Risskov, Denmark. [25]Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, Copenhagen, Denmark.

ONLINE METHODS

The data sets used in these analyses are summarized in **Supplementary Table 1**.

LD score genetic correlation analyses. Genetic correlation quantifies the extent to which two phenotypes share genetic etiology. A genetic correlation of 1 suggests that all influences are shared, whereas a genetic correlation of 0 suggests that the phenotypes are genetically independent. SNP-based genetic correlations have traditionally been estimated using data sets with individual-level genotype and phenotype data. A new method, LD score correlation, allows one to estimate genetic relationships using SNP data when the contributing data sets are siloed[18]. Requiring only GWAS summary statistics, the method uses LD patterns to estimate genetic correlations. The resulting correlations are highly similar to those derived from residual maximum-likelihood analysis (for example, as implemented in GCTA or BOLT-REML[17,29]) and Haseman-Elston regression[30]. LD score regression is implemented in the free and open source software package LDSC (see URLs).

We estimated the genetic association between diagnosed ASDs and traits of social and communication impairment in the general population using LD score correlation. Traits of social and communication impairment were measured using the SCDC in ALSPAC. ALSPAC is a population-based, longitudinal cohort study that initially recruited 14,541 pregnancies in Bristol, UK, with expected dates of delivery between 1 April 1991 and 31 December 1992. All women in the study area with expected delivery dates in that time frame were recruited for participation[19,20]. Ethical approval was obtained from the ALSPAC Law-and-Ethics Committee (IRB00003312) and the local research ethics committees, and written informed consent was provided by all parents. The study website contains details on all available data (see URLs). The GWAS of the SCDC in ALSPAC, from which we obtained the SCDC summary statistics, has already been published[23]. Exome sequencing data from family trios are not available in ALSPAC.

The SCDC is a 12-item parent-rated scale that counts traits of social and communication impairment quantitatively. Each question has the option for a response of 0, 1 or 2; the subsequent range of total scores is 0–24. Individuals with diagnosed ASDs, on average, have very high scores on the SCDC, consistent with the disorders' definitional social and communication impairment[21]. As described in the SCDC GWAS, the SCDC was used at multiple time points in ALSPAC[23]. The SCDC is stable over time[22], and individual scores over time are genetically correlated[23]. To reduce multiple testing, we used the SCDC data for children at 8 years of age, as childhood autistic traits are well studied and have been linked to diagnosed ASDs through twin studies[11,12]. We also focused on the data for children of this age as this approach maximizes SCDC sample size (n = 5,628) and, accordingly, the power of the correlation tests.

The ASD case-control GWAS summary statistics come from the PGC-ASD group and iPSYCH-ASD. The PGC-ASD summary statistics were generated using publically available data from a meta-analysis of 5,305 ASD-diagnosed cases and 5,305 pseudocontrols of European descent (see URLs). The pseudocontrol methodology used by the PGC-ASD group and the PGC GWAS analytical pipeline has been reported on extensively[3,31]. Pseudocontrols are built using the untransmitted alleles from each parent at each locus. The subsequent GWAS is immune to population stratification as the pseudocontrols are ancestrally matched to the cases. LD score correlations have already been estimated using these data, as described in Bulik-Sullivan *et al.*[18]. In **Figure 1**, we highlight a subset of the correlations from that manuscript, specifically those associating PGC-ASD with publically available summary statistics from PGC schizophrenia, PGC bipolar disorder and PGC major depressive disorder (see URLs).

The iPSYCH-ASD data are from a new population-based case-control ASD sample derived from the Danish Neonatal Screening Biobank hosted by Statens Serum Institut, comprising dried bloodspots (Guthrie cards) from all individuals born in Denmark since 1981. The samples can be linked to the Danish register system, including the Danish Psychiatric Central Register. DNA extracted from the bloodspots can be successfully amplified and employed in GWAS[32,33]. The iPSYCH-ASD project aims to genotype all Danish individuals with an available bloodspot and an ASD diagnosis in their medical record (ICD codes F84.0, F84.1, F84.5, F84.8 and F84.9). This study has been approved by the Danish research ethical committee system. This analysis employs the iPSYCH-ASD data generated thus far, specifically the first ten genotyping waves of that collection, which contain 7,783 ASD cases and 11,359 controls. All individuals in the sample were born between 1981 and 2005. Genotyping was performed at the Broad Institute. The data were cleaned and analyzed using the same analysis pipeline described in ref. 31 and other previous PGC GWAS publications.

The GWAS summary statistics used for the secondary genetic correlation analyses are published and publicly available. Summary statistics from a GWAS of child full-scale IQ have been made public by Benyamin *et al.*[34] and were used in Bulik-Sullivan *et al.*[18] to estimate a genetic association between PGC-ASD and IQ in the general population. More than 40% of contributing individuals in the Benyamin GWAS were from the ALSPAC cohort (mean age of 9 years). We used LD score correlation to similarly estimate a genetic association between iPSYCH-ASD and general population full-scale IQ in childhood.

Each of the genetic correlation estimates was obtained using constrained-intercept cross-trait LD score regression, which yields much lower standard errors (~30% lower) than unconstrained-intercept LD score regression[18]. Unlike unconstrained-intercept LD score regression, constrained-intercept LD score regression can give biased estimates of genetic correlation if the two studies have either (i) hidden sample overlap or (ii) shared population stratification. We can rule out both of these concerns in the present study because the PGC-ASD sample is family based (case-pseudocontrol) and does not have case or control overlap with any other PGC analyses. The iPSYCH-ASD and ALSPAC samples do not contain any cases or pseudocontrols that were used in PGC-ASD. In addition, case-pseudocontrol analyses are immune to confounding from population stratification; thus, it is not possible for genetic correlation estimates to be biased by population stratification when at least one of the studies uses a case-pseudocontrol design. As a robustness check, we verified that the point estimates of genetic correlation obtained with unconstrained-intercept LD score regression were similar to the constrained-intercept results presented in **Figure 1** (although the standard errors were higher, as a result of the lower statistical efficiency of unconstrained-intercept LD score regression). These results are presented in **Supplementary Table 3**.

***De novo* variant analyses.** The SSC resource is unique in its combination of detailed phenotypic and genotypic characterizations. The SSC ascertained over 2,800 individuals with ASDs and their nuclear family members, restricting recruitment to families in which no other cases of ASD have been diagnosed out to first cousins. Families were also excluded in the event of intellectual disability in a sibling or a history of parental schizophrenia[24]. To our knowledge, the SSC siblings currently constitute the most deeply phenotyped control sample with available *de novo* variant information. We used these data to examine the relationship between *de novo* variant burden and phenotypic variation in the SSC, in both siblings (n = 1,861) and probands (n = 2,497).

We limited the analysis to *de novo* variant classes that are strongly associated with ASD risk, specifically LoF and DCM variants. LoF mutations include frameshift, splice-site and nonsense mutations. DCM mutations include variants that (i) are predicted to be damaging by PolyPhen-2 and (ii) occur in a gene known to be intolerant of heterozygous missense variation[27,28]. In the SSC, LoF mutations are found in 8.9% of controls and 15.0% of cases; DCM mutations are found in 4.3% of controls and 7.1% of cases. Restricting the classes of *de novo* variants analyzed increases the probability of association with case status as well as with phenotypic variation within cases[26]. Synonymous variants, for example, are associated with neither case status[6] nor phenotypic variation in IQ (P = 0.44) or Vineland score (P = 0.96) in SSC cases. To increase signal, we further filtered observed *de novo* variants on the basis of their presence/absence in the publically available ExAC database (see URLs). The ExAC database contains jointly called exomes from 60,706 adult individuals, recruited for an array of exome sequencing studies. These individual exomes form a reference panel of consistently processed human exonic variation, with particular usefulness for the consideration of rare variants. Just as selection reduces the probability that LoF variants will be seen in genes with low tolerance for functional disruption[27], reference genomes will be less likely to contain variants that create strong risk for reproductively deleterious

phenotypes such as ASDs. In other words, one expects that *de novo* variants not seen in 60,706 reference exomes will be, on average, more deleterious. We considered an SSC *de novo* variant to be recurrent if a variant with the same chromosome, position, reference allele and alternate allele was seen in the ExAC database. After filtering out the variants seen in the ExAC database, the LoF + DCM variant rate was 19.0% in cases and 9.9% in controls.

We have previously associated *de novo* variant burden with variation in case IQ and other measures of case severity, including Vineland score, in the SSC[26]. IQ was not measured in SSC siblings; however, the Vineland was used in a manner consistent with its use in cases. The Vineland assesses overall adaptive functioning, with subscales assessing (i) social ability and (ii) communication ability, as well as (iii) daily living skills. The Vineland is commonly used as one measure of case severity in ASD, although its social and communication subscales are not designed to capture canonical autism-like symptomology. In this analysis, we used Poisson regression, in both cases and controls, to associate Vineland scores with *de novo* variant burden. Sex was controlled for in all *de novo* variant analyses.

29. Loh, P.R. *et al.* Contrasting regional architectures of schizophrenia and other complex diseases using fast variance components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
30. Bulik-Sullivan, B. Relationship between LD score and Haseman-Elston regression. *bioRxiv* doi:10.1101/018283 (20 April 2015).
31. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
32. Børglum, A.D. *et al.* Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. *Mol. Psychiatry* **19**, 325–333 (2014).
33. Hollegaard, M.V. *et al.* Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet.* **12**, 58 (2011).
34. Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic, and associated with *FNBP1L*. *Mol. Psychiatry* **19**, 253–258 (2014).