Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants

F. Kyle Satterstrom1,2,3*, Raymond K. Walters1,2,3, Tarjinder Singh1,2,3, Emilie M. Wigdor1,2,3, Francesco Lescai4,5,6, Ditte Demontis4,5,6, Jack A. Kosmicki1,2,3, Jakob Grove4,5,6,7, Christine Stevens1, Jonas Bybjerg-Grauholm4,8, Marie Bækvad-Hansen4,8, Duncan S. Palmer1,2,3, Julian B. Maller1,2,3, iPSYCH-Broad Consortium9, Merete Nordentoft4,10, Ole Mors4,11, Elise B. Robinson1,2,3,12, David M. Hougaard4,8, Thomas M. Verge4,13,14, Preben Bo Mortensen4,5,15,16, Benjamin M. Neale1,2,3,17, Anders D. Børglum4,5,6*, and Mark J. Daly1,2,3,17,18*

The exome sequences of approximately 8,000 children with autism spectrum disorder (ASD) and/or attention deficit hyperactivity disorder (ADHD) and 5,000 controls were analyzed, finding that individuals with ASD and individuals with ADHD had a similar burden of rare protein-truncating variants in evolutionarily constrained genes, both significantly higher than controls. This motivated a combined analysis across ASD and ADHD, identifying microtubule-associated protein 1A (MAP1A) as a new exome-wide significant gene conferring risk for childhood psychiatric disorders.

ASD and ADHD are substantially heritable1–3, but individuals with psychiatric diagnoses often do not have blood drawn as part of a routine medical procedure, making it difficult to collect cohorts for genetic analysis—particularly for ADHD, which has not previously been the subject of a large-scale sequencing study. To overcome this challenge, we drew upon two Danish national resources: the Danish Neonatal Screening Biobank (DNSB) and the Danish Psychiatric Central Research Register (DPCRR).

As part of the iPSYCH research initiative4, we identified individuals with psychiatric diagnoses using the DPCRR, and we extracted DNA from their archived dried blood samples stored in the DNSB. Individuals were born in Denmark between 1981 and 2005 and were matched to diagnoses of ASD, ADHD, schizophrenia, bipolar disorder, affective disorder and anorexia, as well as intellectual disability (ID), conferred by the end of 2016. We have previously validated the genotyping5 and sequencing6 of archived samples (see Methods), and in this study we exome-sequenced a subset of the DNA samples genotyped in recent common variant analyses of both ASD7 and ADHD6. After quality control, the dataset included 3,962 cases with ASD, 901 cases with both ASD and ADHD, 3,477 cases with ADHD and 5,002 controls without any of the above diagnoses (Table 1).

Studies of de novo variants in ASD have found that the greatest excess of point mutations carried by affected children resides in protein-truncating variants (PTVs; for example, nonsense, frameshift and essential splice-site mutations)7–9. Furthermore, this excess burden is almost exclusively carried by PTVs that are rare in the general population and that occur in likely haploinsufficient genes (that is, probability of being loss-of-function intolerant (pLI) of at least 0.9)7–9. Although we could not call de novo variants in our case-control data, we used these findings to guide our analysis. We defined as ‘rare’ any variant with an allele count no greater than 5 across the combination of our dataset (n = 13,342) with non-Finnish Europeans from the nonpsychiatric exome subset of the Genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org/)(n = 44,779), a total population of 58,121 people. We took special interest in genes with pLI ≥ 0.9, which we termed ‘constrained’.

Results
Rates of constrained rare variation. In samples without ID, we observed a significant excess of constrained rare PTVs (crPTVs) in ASD cases (0.298 per person, P = 1.7 × 10−4 by logistic regression), cases with both ASD and ADHD (0.284 per person, P = 2.5 × 10−4) and ADHD cases (0.279 per person, P = 7.2 × 10−4) compared to

---

1Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 2Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. 4The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark. 5SEQ, Centre for Integrative Sequencing, Aarhus University, Aarhus, Denmark. 6Department of Biomedicine—Human Genetics, Aarhus University, Aarhus, Denmark. 7Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. 8Centre for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. 9Mental Health Services in the Capital Region of Denmark, Mental Health Centre Copenhagen, University of Copenhagen, Copenhagen, Denmark. 10Psychosis Research Unit, Aarhus University Hospital, Risskov, Denmark. 11Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. 12Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, Roskilde, Denmark. 13Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. 14National Centre for Register-based Research, Aarhus University, Aarhus, Denmark. 15Centre for Integrated Register-based Research, Aarhus University, Aarhus, Denmark. 16Department of Medicine, Harvard Medical School, Boston, MA, USA. 17Institute for Molecular Medicine Finland, HI LifE, University of Helsinki, Helsinki, Finland. 18A list of members and affiliations appears in the Supplementary Note. *e-mail: satterst@broadinstitute.org; anders@biomed.au.dk; mjdaly@atgu.mgh.harvard.edu
controls (0.210 per person) (Fig. 1a; see Supplementary Fig. 1a and Supplementary Table 1). Consistent with previous observations, we also observed substantially higher rates of crPTVs in cases with comorbid ID compared to controls (0.404 per person in ASD, \( P = 2.5 \times 10^{-21}; 0.419 \) in ASD + ADHD, \( P = 1.1 \times 10^{-5}; 0.362 \) in ADHD, \( P = 2.3 \times 10^{-7} \) (Fig. 1a; see Supplementary Fig. 1a). By contrast, none of our case categories had a significantly higher burden of rare PTVs in genes with pLI \(< 0.9 \) compared to controls (see Supplementary Fig. 1b). Rates of constrained rare synonymous variation were similar across sample categories (with no case category significantly different from controls), showing that the excess crPTVs in cases did not result from technical differences in variant calling (see Supplementary Fig. 1c). Rates of crPTVs were higher in females than in males across most phenotype groups (see Supplementary Table 1), consistent with a female protective effect\(^{16} \), although differences between the sexes were not significant. Most crPTVs were found in people with precisely one of them (see Fig. 1a; see Supplementary Table 1), consistent with a female protective effect. Rates of constrained rare synonymous variation were similar across sample categories (with no case category significantly different from controls), showing that the excess crPTVs in cases did not result from technical differences in variant calling (see Supplementary Fig. 1c). Rates of crPTVs were higher in females than in males across most phenotype groups (see Supplementary Table 1), consistent with a female protective effect\(^{16} \), although differences between the sexes were not significant. Most crPTVs were found in people with precisely one of them (see Supplementary Fig. 2 and Supplementary Table 2). A similar trend to crPTVs was observed with rare missense variants, although the signal was less pronounced (for example, 0.88 variants per person in ASD cases without ID compared to 0.81 in controls, \( P = 4.1 \times 10^{-3} \) by logistic regression) (see Supplementary...

---

**Table 1 | Phenotype breakdown of samples analyzed in the present study**

| Phenotype group | No diagnoses, no ID | 1 diagnosis, no ID | >1 diagnosis, no ID | >1 diagnosis, with ID | Total |
|-----------------|---------------------|--------------------|--------------------|-----------------------|-------|
| ASD             |                     | 2,430              | 661                | 871                   | 3,962 |
| ASD + ADHD      |                     |                    |                    |                       |       |
| ADHD            |                     | 2,360              | 846                | 271                   | 3,477 |
| Control         | 5,002               |                    |                    |                       | 5,002 |

Samples were matched to diagnoses of ASD, ADHD, schizophrenia, bipolar disorder, affective disorder and anorexia, as well as ID.
Fig. 3, Supplementary Fig. 4a and Supplementary Table 1). Here we considered only missense variants with an MPC score (a measure of the deleteriousness of a missense variant based on a regional model of constraint17) of at least 2. A lower degree of enrichment was observed when considering rare missense variants with MPC <2 (see Supplementary Fig. 4b), with synonymous rates largely comparable across phenotype groups (see Supplementary Fig. 4c).

To compare the results of our case–control study with those previously seen in de novo studies of the Simons Simplex Collection (SSC) and Autism Sequencing Consortium (ASC) datasets10,11,15, we examined genes with three or more published rare de novo protein-truncating variants in ASD. Combining all of our cases with an ASD diagnosis (including those with comorbid ADHD and/or ID), we observed a significantly greater burden of rare PTVs in this set of 14 genes (Table 2; \( P = 1.6 \times 10^{-6} \) by logistic regression, odds ratio (OR) = 6.4, \( n = 4,863 \) ASD cases versus 5,002 controls). The only rare PTVs observed in controls were in lysine demethylase 5B (KDM5B), which acts in a potentially recessive manner18; in the other 13 genes, 37 rare PTVs were observed in cases and none in controls. In addition, when applying our rarity threshold to the SSC + ASC data (see Methods), the rate of crPTVs in the case–control Danish data was similar to the combined rates of published de novo and inherited crPTVs (Fig. 1b).

Having observed similar rates of crPTVs between ASD and ADHD (for example, see Fig. 1a), we decided to further explore the overlap of the two disorders. To rule out the possibility of a common comorbidity driving the signal, the next analyses included only those cases with a single diagnosis (for example, no comorbid ASD + ADHD samples, no ID diagnosis and no diagnoses of schizophrenia, bipolar disorder, affective disorder or anorexia) (\( n = 2,430 \) for ASD and \( n = 2,360 \) for ADHD). As with the more inclusive sample groups, these single-diagnosis ASD cases and ADHD cases had similar burdens of crPTVs overall, and both were significantly greater than controls (Fig. 1c; see synonymous rates in Supplementary Fig. 5b), with synonymous rates largely comparable across phenotype groups (see Supplementary Table 4).

Joint ASD and ADHD analysis. Given the similar crPTV burdens in ASD and ADHD cases, we used a c-alpha test19 to determine whether the sets of constrained genes with rare PTVs were similar or distinct in ASD and ADHD. The c-alpha test can be used to test whether two distributions of rare variants have been selected from the same underlying distribution20. Again considering only cases with a single diagnosis, the test did not find a significant difference between ASD and ADHD, but it did find one when comparing either case group to controls (Table 3; see Supplementary Table 3). This result suggests that the crPTVs in individuals with ASD or ADHD occur not only at similar rates, but also in similar sets of genes. The test did not find a significant difference in any pairwise comparison of ASD cases, ADHD cases and controls when considering constrained rare synonymous variation (Table 3) or rare missense variation (MPC ≥ 2) (see Supplementary Table 4).

The finding that ASD and ADHD had similar burdens of crPTVs occurring in similar genes, and that both were distinct from controls, motivated pooling all of the ASD, ASD + ADHD and ADHD cases \( (n = 8,340) \) for the purposes of gene discovery. To increase our

### Table 2 | Rare PTV counts in genes with three or more published14 rare de novo protein-truncating variants in ASD

| Gene | Published rare de novo PTVs in ASD | Published rare de novo PTVs in unaffected children | Danish rare PTVs: ASD, no ID \( (n = 3,775) \) | Danish rare PTVs: ASD, ID \( (n = 1,088) \) | Danish rare PTVs: ASD, total \( (n = 4,863) \) | Danish rare PTVs: controls \( (n = 5,002) \) |
|------|----------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| CHD8 | 6                                | 0                                             | 1                                             | 1                                             | 2                                             | 0                                             |
| ARID1B | 5                              | 0                                             | 3                                             | 0                                             | 3                                             | 0                                             |
| DYRK1A | 5                              | 0                                             | 0                                             | 3                                             | 3                                             | 0                                             |
| SYNGAP1 | 5                             | 0                                             | 0                                             | 4                                             | 4                                             | 0                                             |
| ADNP | 4                               | 0                                             | 0                                             | 2                                             | 2                                             | 0                                             |
| ANK2 | 4                               | 0                                             | 5                                             | 2                                             | 7                                             | 0                                             |
| DSCAM | 4                               | 0                                             | 1                                             | 0                                             | 1                                             | 0                                             |
| SCN2A | 4                               | 0                                             | 1                                             | 3                                             | 4                                             | 0                                             |
| ASH1L | 3                               | 0                                             | 0                                             | 2                                             | 2                                             | 0                                             |
| CHD2 | 3                               | 0                                             | 0                                             | 1                                             | 1                                             | 0                                             |
| GRIN2B | 3                             | 0                                             | 0                                             | 4                                             | 4                                             | 0                                             |
| KDM5B | 3                               | 2                                             | 7                                             | 1                                             | 8                                             | 8                                             |
| POGZ | 3                               | 0                                             | 0                                             | 3                                             | 3                                             | 0                                             |
| SUV420H1 | 3                           | 0                                             | 1                                             | 1                                             | 0                                             | 0                                             |
| Total, all genes | 55                         | 2                                             | 19                                            | 26                                            | 45                                            | 8                                             |
| OR versus control | –                         | –                                             | 3.1                                           | 15.9                                          | 6.4                                           | –                                             |
| OR ± s.e.m. | –                        | –                                             | 2.1–4.8                                       | 10.4–24.3                                     | 4.4–9.5                                       | –                                             |
| \( P \) | –                              | –                                             | \( 7.5 \times 10^{-3} \)                       | \( 9.1 \times 10^{-11} \)                    | \( 1.6 \times 10^{-6} \)                     | –                                             |

Danish ASD data are from all individuals with an ASD diagnosis (including comorbid ADHD and/or ID, \( n = 4,863 \)) and controls (\( n = 5,002 \)). Danish variants were counted as rare if they had an allele count ≤ 5 across the Danish data and non-Finnish Europeans from the nonpsychiatric exome subset of gnomAD. Published SSC + ASC variants were counted as rare if they had an allele count ≤ 5 across the SSC + ASC data and non-Finnish Europeans from the nonpsychiatric exome subset of gnomAD. P values and ORs are for comparison against controls by logistic regression.
and we used a two-tailed Fisher’s exact test to calculate case versus control rates of synonymous variation in controls than cases as we searched the conservative approach of only considering genes with greater rates of synonymous variation in controls than cases as we searched for genes with greater rates of protein-truncating or missense (MPC ≥ 2) variation in cases than in controls (see Methods).

Among constrained genes, the top result in the PTV analysis was MAP1A, in which we observed 11 rare PTVs in Danish cases (four ASD without ID, five ADHD without ID, one ASD with ID and one ASD + ADHD with ID), none in Danish controls and only four in gnomAD (Table 4; see Supplementary Table 5). With a case versus control P value of 4.11 × 10−7, it survives Bonferroni correction for 17,903 genes and is exome-wide significant. MAP1A is highly expressed in the mammalian brain and is important for the organization of neuronal microtubules; a candidate gene study identified an excess of rare missense variants in MAP1A in ASD and schizophrenia21. Although our case–control study includes inherited variation and does not have the power of a de novo study to isolate high-penetrance PTVs, we observed genes flagged by de novo studies—such as ANKRD11, which is associated with ID22, and SCN2A, which is associated with ASD23—among genes with P < 0.01. We also note RA11, which is associated with Smith–Magenis syndrome23, among our top results. A quantile–quantile plot is shown in Supplementary Fig. 6a, and an analogous plot for synonymous variants (see Supplementary Fig. 6b) shows little inflation. In the analysis based on missense variation (see Supplementary Fig. 6c and Supplementary Tables 5 and 6), no genes passed exome-wide significance.

**Discussion**

In summary, we used DNA from archived bloodspots to conduct an exome sequencing study of ASD and ADHD. To place our study in the context of previous de novo variant studies of ASD, we examined our rare PTVs in the top published ASD genes and found an overwhelming burden in ASD cases compared to controls, suggesting that our study is at least partly tapping into the same signal. We also showed that rates of crPTVs in our ASD cases and controls were consistent with the sum of de novo and transmitted (or untransmitted) crPTV rates previously seen in SSC + ASC data. Next, we...

### Table 3 | C-alpha test results for crPTVs and constrained rare synonymous variants

| Comparison       | Genes | CrPTVs C-alpha P value | Constrained rare synonymous variants Genes C-alpha P value |
|------------------|-------|------------------------|-----------------------------------------------------------|
| ASD versus ADHD  | 932   | 0.93                   | 2,947 0.83                                                |
| ASD versus control| 1,102 | 5.7 × 10−9             | 3,059 0.31                                                |
| ADHD versus control | 1,064 | 1.3 × 10−5             | 3,047 0.93                                                |

ADHD cases (n = 2,430) and ADHD cases (n = 2,360) with only a single diagnosis were tested in pairwise comparisons against each other and against controls (n = 5,002) to determine whether the distributions of genes with crPTVs were significantly different between the phenotype groups. ‘Single’ diagnosis refers to samples with a diagnosis of only ASD or ADHD (that is, no comorbid ASD + ADHD samples, no ID diagnosis and no diagnoses of schizophrenia, bipolar disorder, affective disorder or anorexia). The ‘Genes’ column indicates the number of genes in the comparison with at least one variant.

### Table 4 | Top 15 constrained genes in rare PTV analysis, ranked by two-tailed Fisher’s exact P value, comparing case (n = 8,340) total to combined control + gnomAD (n = 49,781) total variant counts

| Gene      | ASD (n = 3,962) | ASD + ADHD (n = 901) | ADHD (n = 3,477) | Control (n = 5,002) | P (Danish) | gnomAD (n = 44,779) | P (combined OR) | ASD dn | DDD dn |
|-----------|----------------|---------------------|------------------|---------------------|------------|---------------------|-----------------|--------|--------|
| MAP1A     | 5              | 1                   | 5                | 0                   | 9.21 × 10−3 | 4                   | 4.11 × 10−7     | 16.4   | 0      |
| ZNF536    | 2              | 2                   | 3                | 0                   | 3.04 × 10−3 | 0                   | 4.24 × 10−4     | Inf    | 0      |
| SPTBN1    | 1              | 1                   | 1                | 0                   | 1.65 × 10−3 | 2                   | 9.90 × 10−4     | 14.9   | 1      |
| ANKRD11   | 2              | 0                   | 2                | 0                   | 3.04 × 10−1 | 1                   | 1.88 × 10−3     | 23.9   | 2      |
| MAGEL2    | 4              | 0                   | 4                | 0                   | 3.04 × 10−1 | 1                   | 1.88 × 10−3     | 23.9   | 0      |
| RAP1GAP2  | 4              | 0                   | 4                | 2                   | 2.68 × 10−3 | 4                   | 2.10 × 10−3     | 7.2    | 0      |
| SLC2A14   | 3              | 0                   | 3                | 2                   | 7.18 × 10−3 | 3                   | 2.10 × 10−3     | 7.2    | 0      |
| RA11      | 1              | 2                   | 2                | 0                   | 1.65 × 10−3 | 3                   | 2.33 × 10−3     | 10.0   | 1      |
| TNRC6C    | 1              | 2                   | 4                | 0                   | 5.04 × 10−2 | 8                   | 2.78 × 10−4     | 5.2    | 0      |
| GLUL      | 1              | 0                   | 2                | 0                   | 2.97 × 10−1 | 0                   | 2.95 × 10−3     | Inf    | 0      |
| SCN2A     | 3              | 0                   | 3                | 0                   | 2.97 × 10−1 | 0                   | 2.95 × 10−3     | Inf    | 4      |
| STAT5B    | 2              | 0                   | 2                | 0                   | 2.97 × 10−1 | 0                   | 2.95 × 10−3     | Inf    | 0      |
| ZEB2      | 2              | 1                   | 2                | 0                   | 2.97 × 10−1 | 0                   | 2.95 × 10−3     | Inf    | 0      |
| DYNCH1    | 5              | 0                   | 5                | 0                   | 9.01 × 10−2 | 6                   | 3.69 × 10−3     | 6.0    | 0      |
| HSPA12A   | 1              | 1                   | 1                | 1                   | 2.68 × 10−3 | 5                   | 3.69 × 10−3     | 6.0    | 0      |

Cases include all samples with an ASD and/or ADHD diagnosis, regardless of ID status. Controls include all control samples as well as non-Finnish Europeans from the nonpsychiatric exome subset of gnomAD. Only genes with p ≤ 0.9 are shown. P values are also given for comparison of cases to Danish controls (n = 5,002) before combination with gnomAD. /ADHD div/ denotes number of published rare de novo PTVs in ASD (SSC + ASC data, 3,982 probands)24. /DDD div/ denotes number of published rare de novo PTVs in the DDD Study, which examines ID/developmental delay (4,293 probands)25. Note that SCN2A has four PTVs listed in Table 2 but only three listed here because one fell 2 basepairs outside the consensus high-confidence region used when combining with gnomAD (see Methods). Inf, infinite.
observed a similar burden of crPTVs in ASD and ADHD, and this motivated a combined analysis for gene discovery. Using gnomAD as an additional control population, MAP1A was identified as being significantly associated with ASD and ADHD. Because we observe rare MAP1A PTVs in cases both with and without ID—and because the genes near the top of our list are not exclusively those previously identified by de novo studies—our case–control findings may include genes in which protein-truncating variants are relevant to psychiatric cases with milder or more behavioral profiles (and with a contribution from inherited variation), in addition to those characterized by more profound neurodevelopmental symptomatology (and primarily driven by de novo variation).

Genetic connections between ASD and ADHD have been made previously23; for example, twin studies show that traits related to ASD significantly co-occur with traits related to ADHD24, and siblings of children with an ASD diagnosis are more likely to exhibit symptoms of and develop ADHD than the general population25. In the genotype data from our population sample, additional evidence comes from the finding that the two disorders are genetically correlated ($r_g = 0.36$, $P = 1.24 \times 10^{-12}$). This study takes the next step and analyzes exome sequences from individuals with one or both of the disorders, and the similar burden of crPTVs in ASD and ADHD suggests that it is worth investigating whether study designs that have been successful in ASD could also be useful in ADHD. Our results also suggest that cross-disorder rare variant studies could allow investigators to increase the power for gene discovery in a combined analysis, in addition to comparing the contribution of variants across disorders.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of code and data availability are available at [https://doi.org/10.1038/s41593-019-0527-8](https://doi.org/10.1038/s41593-019-0527-8).

Received: 6 April 2018; Accepted: 26 September 2019; Published online: 25 November 2019

References

1. Hallmayer, J. et al. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68**, 1095–1102 (2011).
2. Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
3. Larsson, H., Chang, Z., D’Onofrio, B. M. & Lichtenstein, P. The heritability of clinically diagnosed attention deficit hyperactivity disorder across the lifespan. *Psychol. Med.* **44**, 2223–2229 (2014).
4. 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).
5. 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).
6. Hollegaard, M. V. et al. Archived neonatal dried blood spot samples can be used for accurate whole genome and exome-targeted next-generation sequencing. *Mol. Genet. Metab.* **110**, 65–72 (2013).
7. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
8. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
9. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
10. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
11. Jossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nat. Genet.* **515**, 216–221 (2014).
12. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
13. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
14. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
15. 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).
16. 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).
17. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at bioRxiv [https://doi.org/10.1101/148353](https://doi.org/10.1101/148353) (2017).
18. Martin, H. C. et al. Quantifying the contribution of recessive coding variation to developmental disorders. *Science* **362**, 1161–1164 (2018).
19. Neyman, J. & Scott, E. On the use of c($x$) optimal tests of composite hypothesis. *Bull. Int. Stat. Inst. 41*, 477–497 (1966).
20. Neale, B. M. et al. Testing for an unusual distribution of rare variants. *PLoS Genet.* 7, e1001322 (2011).
21. Myers, R. A. et al. A population genetic approach to mapping neurological disorder genes using deep resequencing. *PLoS Genet.* 7, e1001318 (2011).
22. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
23. Slager, R. E., Newton, T. L., Vlangos, C. N., Finucane, B. & Elsea, S. H. Mutations in RA11 associated with Smith–Magenis syndrome. *Nat. Genet.* **33**, 466–468 (2003).
24. Martin, J. et al. Biological overlap of attention-deficit/hyperactivity disorder and autism spectrum disorder: evidence from copy number variants. *J. Am. Acad. Child Adolesc. Psychiatry* **53**, 761–770 (2014).
25. Ronald, A., Simonoff, E., Kuntsi, J., Asherson, P. & Plomin, R. Evidence for overlapping genetic influences on autistic and ADHD behaviours in a community twin sample. *J. Child Psychol. Psychiatry* **49**, 535–542 (2008).
26. Chien, Y. L. et al. ADHD-related symptoms and attention profiles in the unaffected siblings of probands with autism spectrum disorder: focus on the subtypes of autism and Asperger’s disorder. *Mol. Autism* **8**, 37 (2017).
Methods

Human research. This study was approved by the Regional Scientific Ethics Committee in Denmark and the Danish Data Protection Agency.

gnomAD. All references to gnomAD refer to release 2.1 (beta) of the nonpsychiatric/nonbrain subset that has had samples from psychiatric studies removed (https://gnomad.broadinstitute.org; the dataset in Hail 0.2 format is hosted on the Google Cloud at gs://gnomad-public/release/2.1_beta/ht).

Sample selection. Individuals in the iPSYCH cohort were born in Denmark between 1 May 1981 and 31 December 2005 (ref. 11). Neonatal dried blood samples were stored in the DSNB, which houses samples from nearly all individuals born in Denmark since 1982 (and some from 1981). The iPSYCH initiative considers six primary psychiatric diagnoses—asd, ADHD, schizophrenia, bipolar disorder, affective disorder and anorexia—and individuals were selected for inclusion in the cohort after matching them to psychiatric diagnoses in the DPCRR. At the time of sample selection, diagnoses were those conferred by the end of 2012; in the present study, diagnoses conferred by the end of 2016 were used. ASD cases include individuals with an International Classification of Disease (10th edition; ICD-10 (ref. 12)) diagnosis code of F84.0, F84.1, F84.5, F84.8 or F84.9. ADHD cases include individuals with an F90.0 diagnosis. The ID designation was based on an individual having any diagnosis for ID, including mild, moderate or severe (codes F70–F79).

Sample sequencing and validation. The extraction of DNA from archived DSNB blood samples for whole-genome analysis has been extensively described over the past decade. Publications that form the basis of the present study include articles describing the extraction13, whole-genome amplification14, validation for use in genotyping arrays and validation for use in exome sequencing of DNA from archived DSNB blood samples. Hollecaard et al.15, for example, compared DNA from whole blood samples to DNA from the same individuals extracted from archived dried blood samples of two different ages (3 years and 27 years) and found that the archived samples performed as well as the whole blood samples with regard to error rates in sequencing18.

The DNA used in this study had previously been extracted and whole-genome amplified for use in iPSYCH genotyping studies of common variants in ASD16 and ADHD17. The genotyped iPSYCH cohort consists of over 88,000 samples, and a subset of approximately 20,000 age- and ancestry-matched samples was selected for exome sequencing. A validation study was carried out to confirm that DNA from these samples would generate exome sequences of sufficient quality; Poulsen et al.18 examined variant calls based on DNA from archived DSNB blood samples versus whole blood samples from the same individuals, as well as whole blood samples versus whole blood samples, and found that concordance rates were similar and close to 100%. The Poulsen et al. analysis included samples sequenced at the Broad Institute in Cambridge, MA, USA—which subsequently generated the sequences used in this study—and concluded that whole-genome amplified DNA from archived DSNB samples performed similarly in exome sequencing to DNA from high-quality whole blood samples19.

Following the Poulsen et al. study, sequencing for this study commenced at the Genomics Platform of the Broad Institute using an Illumina Nextera capture kit and an Illumina HiSeq sequencer. Sequencing was carried out in multiple waves, including a smaller pilot wave (Pilot 1) and two larger production waves (Wave 1 and Wave 2). The pilot wave was sequenced close to 100%. The Poulsen et al. analysis included samples sequenced at the Broad Institute in Cambridge, MA, USA—which subsequently generated the sequences used in this study—and concluded that whole-genome amplified DNA from archived DSNB samples performed similarly in exome sequencing to DNA from high-quality whole blood samples19.

After the application of these genotype filters, three call rate filters were used: first the removal of variants with a call rate <90%, then the removal of samples with a call rate <95% (575 samples) and finally the removal of variants with a call rate <95%. Between the sample call rate filter and the final variant call rate filter, one of each pair of related samples was removed using the ibd_prune() function in Hail, defining relatedness as a pi-hat value ≥0.2 (124 samples). Variants remaining in the dataset were annotated with the Variant Effect Predictor20 and one transcript for each variant was selected (prioritizing canonical coding transcripts) to assign a gene and a consequence to each variant. As a final quality control step, samples were removed (505 samples) if they were significantly more non-unique (mean children correction) from the observed mean of number of not-in-gnomAD singletons, based on the probability of drawing the observed number from a Poisson distribution. The purpose of this final step was to remove any of the remaining samples that may have gained noise during the time spent in archive.

After the application of these filters, the dataset contained 16,492 individuals, and the remaining ASD (3,962), ASD + ADHD (901), ADHD (3,477) and control (5,002) samples were selected for use in the present study, while samples with other diagnoses were set aside. ASD cases were 3,005 male and 957 female and had an average birth year of 1992; ASD + ADHD cases were 725 male and 176 female and had an average birth year of 1994; ADHD cases were 2,382 male and 1,095 female and had an average birth year of 1999; controls were 3,592 male and had an average birth year of 1991 (see also Table 1). Allele counts in comparisons with gnomAD—and the combination with it—were calculated within these 13,342 samples.

Statistic P values and OR calculations. For calculating P values and ORs for classes of variants (for example, crPTV rates compared to controls: see Fig. 1a and Supplementary Table 1), logistic regression was performed using the glm function in R (https://cran.r-project.org). Covariates included in the logistic regression model were birth year, sex, the first ten principal components of the genetic data (principal component analysis carried out after dropping non-European samples), number of rare variants, and percentage of exome target covered at a read depth of at least 20, mean read depth at sites within the exome target passing VQSR, number of SNPs (of any population frequency) at sites within the exome target passing VQSR and sequencing wave (one-hot encoded). For Supplementary Fig. 2, the R function chisq.test was used with observed frequencies and Poisson–expected probabilities based on the observed mean, and P values were simulated with 10,000 replicates. For the c-alpha tests, the R package AssotestR (http://cran.r-project.org/web/packages/AssotestR/index.html) was used; 10,000 permutations were run for each pairwise test, and the permutation-based P value was checked to ensure that it was comparable to the reported asymptotic P value (see Supplementary Table 3). For calculating gene-level P values and ORs (for example, see Table 1, Supplementary Table 5 and Supplementary Table 6), a two-tailed Fisher’s exact test was performed using the fisher.test function in R. In all analyses, PTV counts from iPSYCH samples were capped at one per person per gene to correct for the rare situation in which one insertion or deletion event is observed at a single position in a single individual (that is, one insertion or deletion event per person per gene).

Comparison with SSC + ASC. For comparison to our data, we obtained de novo and inherited SSC and ASC data21. Inherited data were obtained directly from J.A. Honer22. To apply the definition of “rare” used in this study for possible variants, in both the de novo and inherited sets of SSC + ASC data were annotated with allele counts from non-Finnish Europeans in the nonpsychiatric exome subset of gnomAD, and variants with an allele count >5 in the combined SSC + ASC + gnomAD group of samples were dropped. Counting the resulting number of rare de novo PTVs per gene produced the list of top genes used in Table 2, the number of HUGO-ASC and HUGO-SSC scores shown in Fig. 1d and Supplementary Fig. 5b and the ASD de novo PTV counts given in Table 4. Integrating de novo crPTV counts with inherited crPTV counts

NATURE NEUROSCIENCE | www.nature.com/natureneuroscience
gave the ‘case’ and ‘control’ crPTV rates constructed for SSC+ASC data in Fig. 1b. Here, case SSC+ASC rates consist of de novo variants in ASD-affected probands (n=3,982) and transmitted variants from parents of probands (n=4,319), while control SSC+ASC rates consist of de novo variants from unaffected children (n=2,078) and untransmitted variants from parents of probands (n=4,319). The Danish ASD data in Fig. 1b are from all children with an ASD diagnosis (with or without ADHD and regardless of ID status, n=4,863), and the Danish control data are the same group of controls (n=5,002) used throughout our analyses.

**Combination with gnomAD.** When combining our data with gnomAD for the purpose of gene discovery, variants were dropped if they fell outside a consensus high-confidence region for the two datasets. This region was defined as the intervals where at least 80% of the samples in both datasets had at least 10x sequencing coverage (based on analysis of bam files for the Danish samples and on coverage summary tables for gnomAD). We considered 17,983 genes overall (after dropping the 59 ACMG genes as mentioned above), and this number was not changed by restricting to the consensus high-confidence region. We then counted the number of rare protein-truncating, missense and synonymous variants by gene. To ensure that the comparison was not biased by differential variation rates between cases (entirely Danish) and controls (mostly gnomAD), all genes with higher rates of rare synonymous variation in cases than controls were excluded (removed 1,615/17,903 or 9.0% of genes). Following this, in the PTV analysis, only genes with higher rates of rare truncating variation in cases than controls were considered (retained 3,182/16,288 or 19.5% of genes). Likewise, in the missense analysis, only genes with higher rates of rare missense variation (MPC≥2) in cases than controls were considered (retained 957/16,288 or 5.9% of genes). As can be seen from these filters, the vast majority of genes had higher rates of variation in controls than in cases, indicating that more rare variants were, on average, being called per sample in gnomAD (potentially due to more liberal quality control thresholds for parameters such as call rate)—a trend that any gene had to overcome to have a greater burden of PTVs or missense variants in cases than controls.

**ID de novo variants.** Table 1 lists the number of published rare de novo PTVs from the Deciphering Developmental Disorders Study (DDD) for each of the top 15 constrained genes in the gene discovery analysis. As none of the published DDD de novo PTVs in these genes had an allele count >5 between the DDD study and the non-Finnish Europeans in the nonpsychiatric exome subset of gnomAD, they were all deemed ‘rare’.

**Quantile–quantile plots.** The PTV quantile–quantile plot (see Supplementary Fig. 6a) displays the 3,182 genes included in the PTV gene discovery analysis, as described above. The synonymous quantile–quantile plot (see Supplementary Fig. 6b) displays all genes with higher rates of synonymous variation in cases than controls (retained 1,615/17,903 or 9.0% of genes). The missense quantile–quantile plot (see Supplementary Fig. 6c) displays the 957 genes included in the missense gene discovery analysis.

**Notes on study design.** All laboratory processing was performed blind to phenotype. Sample selection was of necessity not performed blind to phenotype, but it was performed blind to an individual’s rare variant burden. Sample sizes were set at a number of cases similar to previous useful studies of ASD (for example, ref. 10). To control for downstream batch effects, samples were sequenced in blocks (waves) that included cases and controls matched by birth cohort. The only subjects excluded from this study were filtered due to data quality concerns (described above under ‘Callset quality control’) before analysis. No data points were excluded after starting the analysis. Error bars in bar plots are Poisson s.e.m.; as shown in Supplementary Fig. 2, crPTV distributions did not differ significantly from Poisson expectation. The samples used in the present study are considered reasonable request from F.K.S. (satterst@broadinstitute.org).

**Data availability**

Supplementary data are available as supplementary files to this manuscript (see Supplementary Tables 1, 3 and 5) or at the iPSYCH download page: http://ipsych.au.dk/downloads. For inquiries about more detailed data, contact iPSYCH lead investigator A.D.B. (anders@biomed.au.dk).

**Code availability**

Hail (0.1) and R scripts used to handle and analyze these data are available upon reasonable request from F.K.S. (satterst@broadinstitute.org).

**References**

27. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research.* (World Health Organization, 1993).

28. Hannelius, U. et al. Phenylketonuria screening registry as a resource for population genetic studies. *J. Med. Genet.* **42**, 660 (2005).

29. Hollegaard, M. V. et al. Whole genome amplification and genetic analysis after extraction of proteins from dried blood spots. *Clin. Chem.* **53**, 1161–1162 (2007).

30. Poulsen, J. B. et al. High-quality exome sequencing of whole-genome amplified neonatal dried blood spot DNA. *PLoS One* **11**, e0153253 (2016).

31. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

32. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SFv2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).

33. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

**Acknowledgements**

The iPSYCH project is funded by the Lundbeck Foundation (grant nos. R102-A9118 and R155-2014-1724) and the universities and university hospitals of Aarhus and Copenhagen. The DNSB resource at the Statens Serum Institut was supported by the Novo Nordisk Foundation. Sequencing of iPSYCH samples was supported by grants from the Simons Foundation (grant no. SPAR 311789 to M.J.D.) and the Stanley Foundation. Other support for this study was received from the National Institute of Mental Health (grant nos. U01MH094432-02, U01MH116660-02 and U01MH100229 to M.J.D.). Computational resources for handling and statistical analysis of iPSYCH data on the GenomeDK and ComputeM HPC facilities were provided by, respectively, the Centre for Integrative Sequencing, iSEQ, Aarhus University, Denmark (grant to A.D.B.) and iPSYCH.

**Author contributions**

F.K.S. performed the analysis. R.K.W., T.S., E.M.W., F.L., D.D., I.A.K., J.G., D.S.P. and J.B.M. contributed to the analysis. F.K.S., R.K.W., C.S., J.B.-G., M.B.-H., M.N., O.M., D.M.H., T.M.W., P.B.M., A.D.B. and the iPSYCH–Broad Consortium were involved in sample selection, handling, processing and quality control. M.N., O.M., E.B.R., D.M.H., T.M.W., P.B.M., B.M.N., A.D.B. and M.J.D. were the project core principal investigator group. M.J.D. directed the project. B.M.N. and A.D.B. contributed to project direction. F.K.S. and M.J.D. wrote the manuscript.

**Competing interests**

T.M.W. has acted as advisor and lecturer to the pharmaceutical company H Lundbeck A/S. B.M.N. is a member of the scientific advisory board at Deep Genomics and is a consultant for Camp4 Therapeutics, Takeda Pharmaceuticals and Biogen Inc.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0527-8.

Correspondence and requests for materials should be addressed to F.K.S., A.D.B. or M.J.D.

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

Nature Research 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 Research policies, see Authors & Referees 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**

Raw sequencing data was processed using the Genome Analysis Toolkit (GATK) version 3.4 to produce a VCF version 4.1 variant callset file.

**Data analysis**

After creation of a VCF file, exome sequencing data was handled by Hail version 0.1. Within Hail, variant consequences were assigned using the Variant Effect Predictor (VEP). After quality control in Hail, results were analyzed and plotted in R version 3.2.2. Scripts used are available upon request.

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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Supplementary data are available as supplementary files to this manuscript (Tables S1, S3, and S5) or at the IPSYCH download page: http://ipsych.au.dk/downloads/. For inquiries about more detailed data, contact IPSYCH lead investigator A.D.B. (anders@biomed.au.dk).
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**

Sample sizes were set at a number of cases similar to previous useful studies of ASD. For example, our cohort of 4,863 samples with an ASD diagnosis (after quality control, including 901 ASD+ADHD samples) was of a size similar to the De Rubeis et al. study (2014, Nature 515:209-215, Ref. 10 in the main text, which totaled 3,871 ASD cases across trio-based samples and case-control samples). This suggested that our study could be worthwhile as well, particularly when combined with an additional 3,477 samples with an ADHD diagnosis (after quality control).

**Data exclusions**

To ensure the quality of our samples and the validity of the analysis, samples were dropped from the raw set of sequenced samples if they lacked complete phenotype information, if their imputed sex did not clearly match their reported sex, if they were a duplicate (or monozygotic twin), if they were not putatively European by principal components analysis of their genetic variation, if they were a control (i.e. without a diagnosis of ASD, ADHD, schizophrenia, bipolar disorder, affective disorder, or anorexia) with a diagnosis of intellectual disability, or if they had an estimated levels of contamination or chimeric reads above 5%. In addition, after applying genotype- and variant-level filters (Methods), samples were removed if they had a call rate below 95%, and one of each related pair of samples (defined as a pi-hat above 0.2) was also removed. Finally, to filter noisy samples, samples were removed if they were significantly different (after Bonferroni correction) from the observed mean of number of not-in-gnomAD singletons, based on the probability of drawing the observed number from a Poisson distribution. All sample filters were applied prior to analysis; no data points were excluded after beginning the analysis.

**Replication**

We show in the text that our findings (in addition to their novel aspects) replicate genetic signals identified by previous SSC+ASC studies. Table 2 shows that we have a preponderance of rare truncating variants in autism cases in genes where previous studies identified the most de novo truncating mutations in autism probands. In addition, Figure 1b shows that our rates of rare truncating variants are comparable to rates of de novo + inherited truncating variants in the SSC+ASC data. Within our own data, elevated rates of constrained rare PTVs in cases with autism and/or ADHD as compared to controls were present in each sequencing wave. We therefore conclude that our attempts at replication were successful; our findings both reproduce findings from separate datasets and are also themselves reproducible.

**Randomization**

Allocation of individuals into phenotypic groups was not random but rather relied on whether an individual had received a diagnosis of psychiatric disorder (e.g. autism and/or ADHD and intellectual disability). Selection of controls was random from within birth cohorts matched to the cases. Samples were sequenced in blocks (waves) that included cases and controls matched by birth cohort.

**Blinding**

All laboratory processing was performed blind to phenotype. Sample selection was necessarily not performed blind to phenotype, but it was performed blind to an individual’s rare variant burden.

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 | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| ☒ Involved in the study         | ☒ Involved in the study |
| ☒ Antibodies                    | ☒ ChIP-seq |
| ☒ Eukaryotic cell lines         | ☒ Flow cytometry |
| ☒ Palaeontology                 | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms   |         |
| ☒ Human research participants   |         |
| ☒ Clinical data                 |         |

Human research participants

Policy information about studies involving human research participants

Population characteristics

Samples included people of different sexes and ages and were sequenced in multiple waves. These characteristics were included as covariates (along with principal components of the genetic variation and selected sequencing quality metrics) in logistic regressions comparing rates of different variant types across phenotype groups. Briefly, controls were 3373 male and 1629 female and had an average birth year of 1991; ASD cases were 3005 male and 957 female and had an average birth year of 1992;
| Recruitment | Samples were selected for sequencing from amongst samples that had already been genotyped as part of the iPSYCH study. These samples had initially been selected for genotyping by using the Danish national register of psychiatric diagnoses to identify archived blood samples in the Danish Neonatal Screening Biobank from individuals who would go on to receive diagnoses of disorders such as autism and ADHD. This study design is enabled both by these Danish national resources as well as Danish law. Individuals (or, when they are young, their parents) may opt out having their samples used in research studies, but because the vast majority of samples remain, and because this study does not involve the recruitment of participants, it avoids many of the potential sample selection biases of human subjects research. |
| Ethics oversight | This study was approved by the Regional Scientific Ethics Committee in Denmark and the Danish Data Protection Agency. |

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