# Towards a gene-level map of resilience to genetic variants associated with autism

Thomas Rolland[1*], Freddy Cliquet[1], Richard J.L. Anney[2], Nicolas Traut[1,3], Alexandre Mathieu[1], Guillaume Huguet[4,5], Claire S. Leblond[1], Elise Douard[4,5], Frédérique Amsellem[1,6], Simon Malesys[1], Anna Maruani[1,6], Roberto Toro[1,3], Alan Packer[7], Wendy K. Chung[7,8], Sébastien Jacquemont[4,5], Richard Delorme[1,6], Thomas Bourgeron[1*]

[1] Human Genetics and Cognitive Functions, Institut Pasteur, UMR3571 CNRS, Université de Paris, Paris, France

[2] MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, CF24 4HQ, UK

[3] Center for Research and Interdisciplinarity (CRI), Université Paris Descartes, Paris, France

[4] Department of Pediatrics, Université de Montréal, Montreal, QC, Canada

[5] Centre Hospitalier Universitaire Sainte-Justine Research Center, Montreal, QC, Canada

[6] Department of Child and Adolescent Psychiatry, Assistance Publique-Hôpitaux de Paris, Robert Debré Hospital, Paris, France

[7] Simons Foundation, New York, NY, USA.

[8] Department of Pediatrics, Columbia University Medical Center, New York, NY, USA.

[*] e-mail: thomas.rolland@pasteur.fr , thomas.bourgeron@pasteur.fr

**ABSTRACT**

While over 100 genes are now significantly associated with autism spectrum disorders (ASD), the penetrance of the variants affecting these genes remains poorly understood. Here, we quantified the prevalence of rare loss-of-function (LoF) mutations affecting 156 genes robustly associated with ASD (SPARK genes) using genetic data from more than 10,000 individuals with ASD and 100,000 undiagnosed individuals. We then investigated the clinical, brain imaging and genetic profiles of individuals heterozygous for these rare deleterious variants who were not diagnosed with ASD. These “resilient” individuals, observed in less than 1% of the general population, were equally distributed among males and females, but displayed low polygenic scores for ASD compared to LoF heterozygotes diagnosed with ASD. The interplay between rare and common variants may therefore contribute to the clinical profiles of the individuals carrying LoF in genes associated with ASD.

**MAIN**

In the past 20 years, there has been tremendous progress in identifying genes associated with ASD[1,2], a heterogeneous condition characterized by atypical social communication, as well as restricted or stereotyped interests[3]. The genetic architecture of ASD is complex, ranging from monogenic forms of the disorders (for example caused by a *de novo* variant) to highly polygenic forms, i.e. driven by the additive effect of a large number of common variants, each having a small effect. The SFARI Gene database includes known genes studied in ASD, and among those are 156 listed by SPARK (http://sparkforautism.org) as robustly associated with ASD in multiple independent studies and unrelated individuals (hereafter referred to as SPARK genes, Extended Data Table 1)[4]. In addition to the SPARK genes, often identified through rare mutation screening, genome-wide associations studies (GWAS) have demonstrated that common variants also significantly contribute to ASD[5–7].

Little is known about the prevalence of rare loss-of-function (LoF) variants affecting the function of the SPARK genes in the general population. Nor do we understand the

mechanisms that could underlie the inter-individual differences in clinical manifestations of the carriers[8]. In this study, we provide a gene-level map of the prevalence of LoF variants in the SPARK genes and, in contrast to patient-centric approaches, we investigated non-ASD individuals who seem to be resilient to the presence of such variants. This design could allow the identification of factors protecting against the most severe symptoms associated with such variants[9].

We analyzed whole-exome sequencing (WES) data from 11,067 individuals diagnosed with ASD, 14,988 first-degree relatives of individuals with ASD and 92,977 individuals identified from population cohorts (Extended Data Fig. 1, Extended Data Table 2 and Methods), all of European descent (Extended Data Fig. 2 and Methods). We identified high-confidence rare loss-of-function variants (HC-R-LoF; minor allele frequency, MAF<0.01) in the 156 SPARK genes[4]. Not all SPARK genes act through a loss of function mechanism, but 131 (84%) are considered as intolerant to LoF variants according to previous metrics such as LOEUF (loss-of-function observed/expected upper bound fraction, Extended Data Fig. 3)[10]. Because the impact of a LoF variant might depend on its frequency, its location on the encoded protein and the expression of the corresponding transcripts[10,11], we also selected a more stringent subset of LoF variants (HC-S-LoF) that were (i) singleton, *i.e.* observed only in a single individual or in members of his/her family; (ii) located in an exon present in more than 10% of brain transcripts; and (iii) truncating more than 10% the encoded protein (Extended Data Fig. 4 and Methods).

We observed a significant enrichment of LoF variants in SPARK genes amongst individuals with ASD compared to their relatives and to individuals from the general population (Fig. 1a and Extended Data Table 2). These differences were not observed for synonymous variants (Extended Data Fig. 5). Remarkably, although HC-S-LoF variants in most of the SPARK genes are considered as causative for ASD, we found 0.79% of non-ASD siblings and parents and 0.67% of individuals from the general population carrying these variants. We also observed that LoF variants in non-ASD individuals could affect the

same exons as variants in ASD individuals (Extended Data Fig. 6), suggesting that these mutations should have similar functional consequences[12].

In order to better understand the effect of the LoF variants at the gene level, we estimated for each SPARK gene the attributable risk, i.e. the difference between the fraction of ASD and non-ASD individuals carrying such variants, and the relative risk, i.e. the risk of being diagnosed with ASD when carrying such variants (Methods). Prevalence and risk measures can be visualized and downloaded on https://genetrek.pasteur.fr/. Overall, as expected, LoF variants were more often identified in individuals with ASD compared to the non-ASD individuals (Extended Data Fig. 7), with an average per-gene attributable risk (AR) of 1 in 5,000 individuals with ASD (average AR of 1.9e-04, 95% confidence interval, CI [3.4e-05 – 3.4e-04] for HC-R-LoFs and average AR of 1.8e-04, 95% CI [4.8e-05 - 3e-04] for HC-S-LoFs) (Fig. 1b). Notably, we observed that HC-S-LoFs conferred a higher relative risk (RR) compared to HC-R-LoFs, confirming the utility of taking into account the mutation frequency, the degree of truncation and of the level of exon usage in the transcript before interpreting LoF variants (p=0.035, Extended Data Fig. 4)[11].

Several genes such as *CHD8*, *DYRK1A* and *SCN2A* displayed high AR indicating that they were among the most frequently mutated genes for ASD (Fig. 1b and Extended Data Fig. 8). However, these genes confer different RR (e.g. HC-S-LoF $RR_{CHD8}$=13; $RR_{DYRK1A}$=39.6, $RR_{SCN2A}$=infinite). *SCN2A* is among 26 SPARK genes such as *GRIN2B* and *SYNGAP1* for which we could not find any carrier of HC-S-LoFs among the 107,965 non-ASD individuals, suggesting that LoF variants in these genes have very large effect sizes (fully penetrant) and are exclusively found *de novo*[13]. In contrast, for 78 SPARK genes including *CHD8*, *DYRK1A* and *POGZ*, we could identify at least one carrier of HC-S-LoF variants among the non-ASD individuals, suggesting incomplete penetrance for ASD diagnosis (Fig. 1b, Extended Data Table 3). Overall, the gene-level RR was highly correlated to other measures of LoF intolerance[10], yet some genes such as *PTEN* and *SHANK2* conferred RR above 5 while being relatively tolerant to LoF mutation according to the LOEUF scale (Extended Data Fig. 9). This discrepancy between RR and gene metrics has been

previously observed, supporting caution in application of specific cutoffs[14], and underlines the need to investigate less penetrant variations to complete the map of genes associated to ASD[15].

To further investigate the relationship between biological functions and the RR for ASD, we studied the expression level of SPARK genes in different human brain regions and at different developmental periods, and found a correlation between gene-level RR and expression in the early fetal period (8-12 post-conception weeks) of cortex development (Fig. 1c, Extended Data Fig. 10, Extended Data Table 4 and Methods)[16]. We found no significant correlation between known ASD-associated pathways and higher RR of HC-S-LoFs (Extended Data Fig. 10 and Extended Data Table 5).

While ASD displays a strong gender bias, with 1 girl for 3-8 boys, in part due to current scales used for ASD diagnosis being less well suited to girls than to boys[9,17], we observed among non-ASD individuals no difference in sex ratio between LoF carriers and non-carriers (Extended Data Fig. 11). Of note, ASD individuals carrying LoFs were significantly enriched in females (OR=1.7, p=0.0025) as previously reported[18,19]. We next explored if the non-ASD individuals carrying LoF variants displayed partial phenotypes at the global functioning, cognitive or brain levels. In the SSC and SPARK families, none of the siblings carrying LoF variants in SPARK genes had total scores on the social communication questionnaire (SCQ) above the suggested cut-off of 15, confirming an absence of autistic traits (Fig. 1d and Extended Data Fig. 11). In the UK-Biobank cohort, we investigated if the non-ASD carriers of LoF variants displayed differences in cognitive and socio-economical features compared to non-carriers. We observed no difference in mental distress history between LoF variant carriers and non-carriers (Extended Data Fig. 11), but remarkably they displayed lower qualification levels, lower incomes, lower fluid intelligence scores and higher material deprivation (Fig. 2a and Extended Data Fig. 12). These differences increased when considering genes with high RR (Fig. 2b), indicating a gradient of severity of LoF variants in the general population as previously reported for large copy-number variants[20–22].

We also investigated the brain structural anatomy of the non-ASD LoF variants carriers by analyzing the brain magnetic resonance imaging (MRI) data from 11,890 individuals from UK-Biobank (Methods). We identified 67 carriers of HC-S-LoFs with MRI data and compared the total brain volume and volumes of 13 brain substructures to the non-carriers. Overall, although they carry LoF variants in SPARK genes, the distribution of brain volumes of the carriers was not significantly different from the non-carriers (Extended Data Fig. 13) at all thresholds for RR (Extended Data Table 6). To move to the gene level and to study biological functions that could be associated to brain volume variations, we clustered the genes based on the average volume differences between corresponding HC-S-LoF carriers and non-carriers (Fig. 2c and Extended Data Fig. 14). We observed that a first cluster of 25 genes associated with smaller brain volumes was notably enriched in genes related to modulation of excitatory post-synaptic potential (OR=7, p=0.027), regulation of gene expression (OR=2.2, p=0.018), and regulation of macromolecule metabolic process (OR=2.1, p=0.028) (Fig. 2d). The second cluster of 13 genes associated with larger brain volumes included *PTEN*, known to be linked with macrocephaly[23], and was enriched in genes involved in regulation of synaptic transmission (OR=10.1, p=0.001).

Finally, to test if the genetic background could influence the penetrance of the mutation[24], we measured the genome-wide polygenic score for ASD (ASD-PGS)[7,25] for 52,630 individuals from the SSC, SPARK and UK-Biobank cohorts (Methods). We then compared the fraction of individuals from the different groups in the top tercile of the ASD-PGS scores distribution. Among non-carriers, we observed that ASD individuals were 1.3 and 2.1 more likely to be in the top tercile of the distribution than non-ASD siblings and UK-Biobank individuals, respectively (Fig. 2e). In contrast, among HC-S-LoF carriers, the odds ratio increased to 3 and 3.4 between ASD individuals and non-ASD siblings or UK-Biobank control individuals, respectively. The same trend was observed among HC-R-LoFs (Fig. 2e), with ASD individuals up to 5 and 3.7 more likely to be in the top tercile of the distribution compared to non-ASD siblings and UK-Biobank control individuals (p=0.039 and p=2e-04, respectively). Interestingly, we observed no enrichment of non-ASD siblings and UK-Biobank

individuals carrying LoF variants among individuals with low ASD-PGS values (Extended Data Fig. 15).

In summary, by systematically analyzing WES data of more than 10,000 ASD and 100,000 non-ASD individuals, our work provided an estimate of the attributable and relative risks associated to each of the curated SPARK genes that are currently considered as the highest-confidence genes for ASD diagnosis. The genes with the highest risk in our meta-analysis were those with high intolerance to LoF variants and repeatedly identified as affected by *de novo* mutations in independent genetic studies of ASD, including *DYRK1A*, *GRIN2B*, *SCN2A* and *SYNGAP1*. In contrast, for some SPARK genes such as *SHANK2* and *PTEN*, LoF variants might be passed through generations and therefore considered to be tolerant to LoF mutations according to the LOEUF score[10].

Interestingly, in some cases LoF variants are not necessarily sufficient for a clinical diagnosis of ASD, but they might however influence the global functioning of the carriers as indicated by some socio-economic metrics. We also observed that some variants could have an effect on brain development, but under a certain threshold for clinical diagnosis, supporting a complex interplay between gene-level variations, modification of brain structure and clinical outcome[23,26]. Although girls were predicted to be more “resilient” to ASD, we did not observe overall differences in sex-ratio amongst non-ASD carriers of LoF variants affecting the SPARK genes. This did not exclude that for specific genes or pathways, the effect of gender could be more important. For example, inherited mutations in autosomal genes such as *SHANK1* or *CHD8* seemed to be more transmitted by mothers and lead to ASD preferentially or exclusively in males[26,27].

Our observation that high polygenic scores for ASD increased the probability of having an ASD diagnostic in the carriers of the LoF variants suggests again that common and rare variants act additively on the probability to be diagnosed with ASD[5]. Integration of additional polygenic scores and data related to expression levels (expression quantitative trait loci, eQTL) could provide a better estimate of the developmental trajectory of the

carriers[24,28,29] and enhance our understanding of the neurological and behavioral manifestations[6,30].

The mechanisms of resilience to deleterious mutations will be however difficult to identify since proteins encoded by ASD-associated genes have diverse functions in complex cellular systems involving synapses, chromatin remodeling, transcription and translation[3]. Hence, the impact of a mutation on the function of the encoded protein may be compensated by distinct mechanisms involving other genes with partial functional redundancy or alternative pathways.

In summary, our results provide a proof-of-concept that investigating resilience to deleterious genetic mutations on large genetic and phenotypic datasets of individuals could inform on clinical outcomes. Future integrative analyses of genetic, epigenetic, transcriptomic and proteomic aspects of brain-related biological functions will shed light on the different biological processes underlying ASD and should open the path towards the identification of factors that can protect individuals from severe outcomes.

## FIGURE LEGENDS

**Figure 1. Risk and resilience to ASD at the mutation and gene level.** (**a**) Fraction of individuals carrying high-confidence rare (HC-R-LoF, top) and stringent (HC-S-LoF, bottom) loss-of-function variants in SPARK genes in each cohort, stratified by status and family relationship. Error bars correspond to standard errors of the proportions. Odds ratios and p-values from two-sided Fisher exact tests comparing fractions of carriers amongst ASD and control individuals. SSC: Simons Simplex Collection cohort, SPARK: Simons Powering Autism Research for Knowledge cohort, DBS: Danish Blood Spot (iPSYCH) cohort. (**b**) Relative risk and attributable risk of HC-S-LoF variants for each SPARK gene for which at least one variant was found in an ASD individual. The 95% confidence intervals (CIs) are shown. Inf.: Infinite value due to no variants identified among non-ASD individuals. (**c**) Correlation between relative risk of HC-S-LoF variants and gene expression in four brain

regions (top to bottom: somatosensory, motor and prefrontal cortex; parietal, auditory, visual and temporal cortex; striatum, hippocampus and amygdala; mediodorsal nucleus of the thalamus and cerebellar cortex) and developmental periods for the studied SPARK genes. Correlations and p-values measured by one-sided Kendall correlation tests of correlation between relative risk and gene expression (nominally significant p-values are shown, *nominal $p<0.05$). SPARK genes for which none of the non-ASD individuals were identified as carrying HC-S-LoF variants were removed for the analysis, and for SPARK genes for which none of the ASD individuals were identified as carrying HC-S-LoF variants, we artificially set the number of carrying ASD individuals to 1 so that relative risk measures were not infinite (Methods). (**d**) Distribution of SCQ scores of ASD individuals and non-ASD siblings carrying and not carrying HC-S-LoF variants.

**Figure 2. Socio-economical profile, brain structure and genetic background of non-ASD LoF carriers.** (**a**) Fraction of HC-S-LoF carriers and non-carriers in each qualification level (from white to grey: CSEs or equivalent, O levels/GCSEs or equivalent, NVQ or HND or HNC or equivalent, A levels/AS levels or equivalent, College or University degree) and income level (from white to grey: less than 18,000, 18,000 to 30,999, 31,000 to 51,999, 52,000 to 100,000, greater than 100,000 pounds; p-values from chi-square tests), and distribution of Townsend Deprivation Index and fluid intelligence score (means and standard deviations are shown, p-values from one-sided Wilcoxon rank tests) for HC-S-LoF carriers and non-carriers among UK-Biobank control individuals. (**b**) Regression coefficients associated to HC-S-LoFs for fluid intelligence score and Townsend index of deprivation (linear regressions), and income and qualification levels (ordinal regressions) in the UK-Biobank cohort (95% CIs are shown, Methods). The direction of effect is adjusted for Townsend index so that higher material deprivation is indicated with a negative sign. P-values for the coefficients associated to HC-S-LoF presence are adjusted for multiple testing using the Benjamini and Hochberg control for false discovery rate. All tests were performed when considering genes over increasing thresholds of relative risk, and the number of

carriers is indicated. (**c**) Heatmap of the difference in volume in different brain regions between carriers and non-carriers of HC-S-LoFs among UK-Biobank individuals. Clustering of genes using a hierarchical clustering method with two clusters. All individuals carrying HC-S-LoFs were considered in this analysis. (**d**) Functional enrichment for each gene cluster is measured using two-sided Fisher exact tests and p-values are adjusted using a stochastic approach (Methods). (**e**) Fraction of HC-S-LoF and HC-R-LoF carriers and non-carriers in the top tercile of the ASD-PGS score distribution, stratified by cohort type (children from ASD families and control individuals) and status. For the SSC and SPARK cohorts, only inherited variants were considered. Odds ratios and p-values were measured using two-sided Fisher exact tests (95% CIs are shown) and p-values were adjusted for multiple testing using the Benjamini and Hochberg control for false discovery rate. All tests were performed for HC-S-LoFs (top) and for HC-R-LoFs (bottom) considering genes over increasing thresholds of relative risk, and the number of variant carriers is indicated.

## REFERENCES

1. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).

2. Feliciano, P. *et al.* Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *Npj Genomic Med.* **4**, 1–14 (2019).

3. Bourgeron, T. From the genetic architecture to synaptic plasticity in autism spectrum disorder. *Nat. Rev. Neurosci.* **16**, 551–563 (2015).

4. Feliciano, P. *et al.* SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).

5. Weiner, D. J. *et al.* Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).

6. Castel, S. E. *et al.* Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).

7. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

8. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34**, 531–538 (2016).

9. Szatmari, P. Risk and resilience in autism spectrum disorder: a missed translational opportunity? *Dev. Med. Child Neurol.* **60**, 225–229 (2018).

10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

11. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

12. Chiang, A. H., Chang, J., Wang, J. & Vitkup, D. Exons as units of phenotypic impact for truncating mutations in autism. *Mol. Psychiatry* 1–11 (2020).

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. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).

15. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).

16. Lin, G. N. *et al.* Spatiotemporal 16p11.2 Protein Network Implicates Cortical Late Mid-Fetal Brain Development and KCTD13-Cul3-RhoA Pathway in Psychiatric Diseases. *Neuron* **85**, 742–754 (2015).

17. Halladay, A. K. *et al.* Sex and gender differences in autism spectrum disorder: summarizing evidence gaps and identifying emerging areas of priority. *Mol. Autism* **6**, 36 (2015).

18. Werling, D. M. & Geschwind, D. H. Sex differences in autism spectrum disorders. *Curr. Opin. Neurol.* **26**, 146–153 (2013).

19. 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).

20. Kendall, K. M. *et al.* Cognitive performance and functional outcomes of carriers of pathogenic copy number variants: analysis of the UK Biobank. *Br. J. Psychiatry* **214**, 297–304 (2019).

21. Huguet, G. *et al.* Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. *JAMA Psychiatry* **75**, 447–457 (2018).

22. Chawner, S. J. R. A. *et al.* A Genetics-First Approach to Dissecting the Heterogeneity of Autism: Phenotypic Comparison of Autism Risk Copy Number Variants. *Am. J. Psychiatry* **178**, 77–86 (2021).

23. Butler, M. G. *et al.* Subset of individuals with autism spectrum disorders and extreme macrocephaly associated with germline PTEN tumour suppressor gene mutations. *J. Med. Genet.* **42**, 318–321 (2005).

24. Davies, R. W. *et al.* Using common genetic variation to examine phenotypic expression and risk prediction in 22q11.2 deletion syndrome. *Nat. Med.* **26**, 1912–1918 (2020).

25. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

26. Bernier, R. *et al.* Disruptive CHD8 Mutations Define a Subtype of Autism Early in Development. *Cell* **158**, 263–276 (2014).

27. Sato, D. *et al.* SHANK1 Deletions in Males with Autism Spectrum Disorder. *Am. J. Hum. Genet.* **90**, 879–887 (2012).

28. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161–165 (2005).

29. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).

30. Hartman, J. L., Garvik, B. & Hartwell, L. Principles for the Buffering of Genetic Variation. *Science* **291**, 1001–1004 (2001).

## METHODS

### Whole exome sequences

We downloaded the GRCh37 aligned BAM files of 8,960 SSC participants from SFARI Base (https://sfari.org/sfari-base). We then called the variants using GATK 3.8 following the BROAD Institute Best Practices[31], and mapped all variants to the GRCh38 human genome version. Three families were filtered out due to a high number of erroneous variant calls (12958, 14572 and 11037). We downloaded the preprocessed GRCh38-based pVCF files of 27,287 SPARK participants from SFARI Base. All functional-equivalent (FE) GRCh38-based gVCF files for 49,972 UK-Biobank participants were downloaded from the UK-Biobank database (projects 51869 and 18584). Variant calling was completed using GATK GenotypeGVCFs[31]. All variants from SSC, SPARK and UK-Biobank cohorts were filtered for call rate > 0.9, genotype quality >= 30, depth > 20, allelic fraction >= 0.25 (and <= 0.75 for autosomal variants). Tabular lists of variants from the Autism Sequencing Consortium Danish Blood Spot individuals were downloaded from the ASC website (https://asc.broadinstitute.org), and mapped to the GRCh38 human genome version (using chain file hg19toHg38.over.chain.gz).

We used VEP[32] (using Ensembl 91) to annotate the variants. Non-neuro (individuals that are not cases of a few particular neurological disorders) non-Finnish European population frequencies were extracted using gnomAD exomes r2.1.1[10]. Variants with a MAF > 1%, present in > 1% of each cohort or affecting genes that were recurrently found mutated across different individuals in different families (*MUC4*, *MUC12*, *HLA-A*, *HLA-B*, *HYDIN*, *TTN*, *PAX5*, *OR2T10*, *MYH4*) were filtered out. We used Loftee[10] to filter low-confidence variants or variants corresponding to ancestral alleles, as well as variants annotated with any flag by Loftee.

We focused on coding exons of 156 genes that are known to be associated to autism (SPARK gene list, downloaded in June 2020, see Extended Data Table 1)[4] and LoF variants affecting these genes were visually validated with IGV[33] on BAM/CRAM files for SSC,

SPARK and UK-Biobank cohorts. Variants in *MECP2* and *PCDH19*, located on the X chromosome and known for their association with ASD among female individuals[34,35], were considered only when identified in female individuals. Variants in *AFF2*, *ARHGEF9*, *ARX*, *ATRX*, *CASK*, *CDKL5*, *DDX3X*, *FMR1*, *HNRNPH2*, *IQSEC2*, *NEXMIF*, *NLGN3*, *PTCHD1*, *SLC9A6* and *UPF3B*, which are also located on the X chromosome, were considered only when identified in male individuals, as they may contribute to ASD when the single copy is mutated.

**Relative position on encoded protein and pext score**

We annotated the relative position of the variants on the encoded protein using the Loftee coding sequence (CDS) position when available or VEP CDS position otherwise, and the CDS size for each transcript from BioMart[36]. To measure exon usage in different isoforms of each gene within brain tissues, we downloaded the base-level pext score from the gnomAD website (https://gnomad.broadinstitute.org)[11]. We averaged the pext measures from 13 brain tissues (amygdala, anterior cingulate cortex BA24, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex BA9, hippocampus, hypothalamus, nucleus accumbens basal ganglia, putamen basal ganglia, spinal cord, substantia nigra). For splice-site variants, we measured the relative position and pext score based on the closest coding exon (position of the variant +/- 3 bp).

**Attributable risk and relative risk**

The attributable risk (AR) and relative risk (RR) were measured to estimate the strength of the association between outcome (ASD diagnostic) and genetic risk factors (carrying a LoF variant). These were calculated as follows, where AE is the number of ASD individuals carrying LoF, AN is the number of ASD individuals not carrying LoF, UE is the number of control individuals carrying LoF, and UN is the number of control individuals not carrying LoF.

Attributable risk:

$$(1)\ \mathrm{AR} = \frac{AE}{AE+AN} - \frac{\mathrm{UE}}{UE+UN} \text{ and}$$

$$(2)\ \mathrm{SE(AR)} = \sqrt{\frac{AE*AN}{(AE+AN)^3} + \frac{\mathrm{UE*UN}}{(UE+UN)^3}} \text{ and}$$

$$(3)\ \mathrm{CI}_{1-\alpha}(AR) = AR \pm SE(AR) * z_\alpha$$

<u>Relative risk:</u>

$$(4)\ \mathrm{RR} = \frac{\frac{AE}{AE+AN}}{\frac{UE}{UE+UN}}$$

$$(5)\ \mathrm{SE(log(RR))} = \sqrt{\frac{AE}{AE*(AE+AN)} + \frac{\mathrm{UE}}{UE*(UE+UN)}}$$

$$(6)\ \mathrm{CI}_{1-\alpha}(\log(RR)) = \log(RR) \pm SE(\log(RR)) * z_\alpha$$

The level of significance $\alpha$ was set to 0.05 to measure the 95% confidence intervals (CI). The confidence interval around the RR was measured as exponentiated confidence intervals around the log RR.

When indicated in the corresponding analyses of correlation involving relative risk, we artificially set the number of control individuals to 1 (UE=1 and UN=UN-1) before calculating attributable risk and relative risk to avoid infinite values of relative risk.

**Alternative deleteriousness scores**

The LOEUF (loss-of-function observed/expected upper bound fraction), pLI (probability of loss-of-function intolerance) and missense z-score metrics were extracted from the gnomAD website (https://gnomad.broadinstitute.org)[11]. The TADA (transmission and de novo association) scores were extracted from a previous study[13].

**SNP arrays**

For the SSC cohort, the GRCh36-based SNP array data for the 3 different technologies (Illumina Omni1Mv1, n = 1,354, Omni1Mv3, n = 4,626, and Omni2.5, n = 4,240) were downloaded from the SFARI Base (https://sfari.org/sfari-base). Arrays from each technology

was mapped onto the GRCh37 human genome version separately. We downloaded the preprocessed GRCh37-based genotyping files of 27,099 SPARK participants from the SFARI Base, and nine families were withdrawn from the original dataset (SF0040498, SF0061696, SF0071614, SF0075822, SF0036255, SF0013311, SF0011132, SF0027703 and SF0110914). SSC and SPARK genotyping files were filtered from ambiguous SNPs (A/T & G/C SNPs if MAF > 0.4, SNPs with differing alleles, SNPs with > 0.2 allele frequency difference, SNPs not in reference panel) and imputed on the HRC panel version r1.1[37] on the Michigan servers with default parameters[38]. GRCh37-based imputed genotyping files for 49,972 UK-Biobank individuals were downloaded from the UK-Biobank database (projects 51869 and 18584), 3,395 individuals were removed because the corresponding SNP arrays did not pass our quality control, 149 individuals were withdrawn from the original dataset, and 38 individuals were removed because they reported ASD-related symptoms (based on ICD10-F84 index or the autism diagnostic questionnaire). After imputation we kept only variants with a r2>=0.8 and merged the 3 different SNP arrays technologies from the SSC cohort keeping only SNPs shared between all 3 technologies.

**Admixture**

For SSC, SPARK and UK-Biobank cohorts, an admixture model was calculated separately using Admixture[39] for 10,219, 26,961 and 46,464 individuals with SNP genotyping and WES data, respectively. The 1000genome sequencing data of 2,504 individuals was used as a reference group of individuals of known ancestry[40] for SPARK and UK-Biobank cohorts, and the HapMap genotyping data of 1,184 individuals for the SSC cohort[41]. We performed admixture analyses on 1 to 8 clusters, and selected 4 clusters for separating the individuals by ancestry, corresponding to the first inflection point for the cross-validation error in the UK-Biobank admixture (Extended Data Fig. 2) and matching superpopulations from the 1000genomes project (African, East Asian, South Asian, European/American). We used a fraction of each individual's SNPs predicted as European ancestry threshold of >= 80% to

define individuals as being of European ancestry, resulting in 6,153, 15,091 and 42,984 individuals in SSC, SPARK and UK-Biobank cohorts, respectively.

**Relatedness among UK-Biobank participants**

We used the Somalier software with default parameters (https://github.com/brentp/somalier) to estimate the relatedness among the 49,972 UK-Biobank participants[42]. For each individual, we extracted the list of variants corresponding to the list of GRCh38-based variants provided by Somalier. We used the identity-by-state (IBS) measures ibs0 < 500 and ibs2 > 8000 to identify individuals with a high level of relatedness corresponding to sibling or parental relationships, and carrying the same HC-R-LoFs (Extended Data Fig. 14).

**ASD polygenic score computation**

SSC, SPARK and UK-Biobank imputed genotyping data were merged with PLINK 1.9[43], filtering out variants absent from less than 1% of each cohort (geno001 parameter). We first performed a PCA analysis using PLINK 2.0 to control for population structure. The first 4 components were used as covariables in PRSice-2 to compute the Genome-wide Polygenic Score (PGS)[44]. The PGS for autism was computed by using the GWAS summary statistics from the Integrative Psychiatric Research Consortium (iPsych) and the Psychiatric Genomics Consortium (PGC)[7]. To exclude overlap in participants from the test and discovery data in the PGS analysis, the GWAS meta-analysis summary statistics reported[7] were recalculated with the SSC data excluded. The resultant PGS were calculated on common (MAF>10%) LD independent SNPs, LD independent SNPs were defined using a clumping window of 500 kb and r2<0.1. The PGS values used in this study were based on a standard p-value threshold of pT = 0.05, using 11,451 variants for a fit R2 of 0.011 and p-value of 1.2e-107, and were z-scored.

**Developmental brain gene expression**

The developmental brain transcriptome data from 42 specimen and up to 16 brain structures was downloaded from the Allen Brain Atlas BrainSpan database (https://www.brainspan.org/). Only expression RPKM values (Reads Per Kilobase of exon model per Million mapped reads) above 1 were considered for expression analysis. Values for each gene were averaged across four brain regions and eight developmental periods as previously described[16].

**Psychiatric, cognitive and functioning feature**

The Social Communication Questionnaire results for SSC and SPARK cohorts were downloaded from SFARI Base (https://sfari.org/sfari-base) and were available for 1,681 and 3,735 probands and 1,225 and 1,863 non-ASD siblings, respectively.

For the UK-Biobank individuals, age when attending assessment center and genetic or declared sex were available for all 42,984 control European individuals. We measured mental distress using the responses to the corresponding online questionnaire ("Have you ever sought or received professional help for mental distress", Yes=1, No=2) available for 19,378 individuals. The Fluid Intelligence Test of reasoning and problem solving was completed by 41,940 individuals. We used the highest qualification an individual had achieved (e.g. university/college degree, A-levels), excluded participants with only 'other professional qualifications' and those who did not provide an answer to this question, retaining data for 34,640 individuals. Annual income was categorized by the UK Biobank in five bands (<£18,000, £18,000–30,999, £31,000–51,999, £52,000– 100,000 and >£100,000), and was available for 37,746 participants. The Townsend Deprivation index of social deprivation was available on 42,940 individuals.

**Brain MRI imaging**

Freesurfer determined volumes were downloaded from the UK-Biobank database for 12,026 European individuals with WES data available. Intra-cranial, total brain, cerebral cortex, cerebral white matter, subcortical structures, corpus callosum and cerebellum grey and white

matter volumes were regressed on sex, age (fourth polynomial degree, poly R function), imaging center and whether MRI T2 FLAIR was used on top of MRI T1D when running Freesurfer. Residuals of the regression (lm R function) were then z-scored. The hierarchical clustering was performed using Euclidean distances (dist R function) and Ward's criterion (hclust R function).

**Statistical analyses**

Most of the statistical analyses in this work were performed using statistical test implementations from the R package[45].

For developmental brain transcriptome analyses, relative risk measures of each SPARK gene were log10 transformed and correlated to log10-transformed expression values in each period/region of brain development (glm R function).

For fluid intelligence score and Townsend Deprivation index, we used linear regression analyses (glm R function) with the score as the dependent variable. We used LoF carrier status as the independent variable and included as covariates sex and age at the time of assessment. Results are presented as unstandardised regression coefficients. Qualifications and household income were analyzed in ordinal regression analyses (polr R function), and results are expressed as the exponential of the odds of carriers being in a different band minus 1 (e.g. in a lower income bracket). For ease of interpretation, the directions of the Townsend Deprivation index effects were adjusted in Fig. 2b, so that a negative sign always implies worse outcome, e.g. higher Townsend Deprivation index. To evaluate the significance of results, we used the Benjamini–Hochberg false discovery rate (FDR) method for p-value correction (p.adjust R function).

For gene ontology term enrichment analyses, we used the funcassociate tool that measured enrichment across the whole gene ontology and provided p-values adjusted for multiple testing based on a stochastic approach[46].

**METHODS REFERENCES**

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. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

33. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

34. Amir, R. E. *et al.* Rett syndrome is caused by mutations in X-linked MECP2 , encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23**, 185–188 (1999).

35. Dibbens, L. M. *et al.* X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment. *Nat. Genet.* **40**, 776–781 (2008).

36. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).

37. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

38. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

39. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

40. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

41. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

42. Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).

43. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

44. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software.

*Bioinformatics* **31**, 1466–1468 (2015).

45. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017).

46. Berriz, G. F., Beaver, J. E., Cenik, C., Tasan, M. & Roth, F. P. Next generation software for functional trend analysis. *Bioinformatics* **25**, 3043–3044 (2009).

## DATA AVAILABILITY

Whole-exome and SNP genotyping data from the SSC and SPARK cohorts can be obtained by applying at SFARI Base (https://www.sfari.org/resource/sfari-base/). The UK-Biobank whole-exome, SNP genotyping and brain imaging data can be obtained by applying at the UK-Biobank database (https://www.ukbiobank.ac.uk/). The human neurodevelopmental transcriptome dataset is available on the BrainSpan database (http://www.brainspan.org). Functional annotations can be obtained from SynGO (https://syngoportal.org/) and Gene Ontology (http://current.geneontology.org/annotations/goa_human.gaf.gz). Electronic health records and healthcare claims data used in the present study for the UK-Biobank individuals are not publicly available due to patient privacy concerns. Prevalence and risk measures can be visualized and downloaded on https://genetrek.pasteur.fr/.

## CODE AVAILABILITY

Code used to implement the post-processing analyses in this paper is available at https://github.com/thomas-rolland/asd-resilience.

## ACKNOWLEDGEMENTS

This research has been conducted using the Simons Simplex Collection and Simons Powering Autism Research for Knowledge from the Simons Foundation Autism Research Initiative, and the UK-Biobank cohort (projects 51869 and 18584). This work was supported by a grant from SFARI (#: 240059, TB). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, at the participating Simons Searchlight

sites, the Simons Searchlight Consortium, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to SNP arrays, WES and phenotypic data on SFARI Base. Approved researchers can obtain the SSC population dataset and the Simons Searchlight population dataset described in this study by applying at https://base.sfari.org. The authors would like to thank the members of the Human Genetics and Cognitive Functions lab for helpful discussions. This work was funded by Institut Pasteur, the Bettencourt-Schueller Foundation, Université de Paris, the Conny-Maeva Charitable Foundation, the Cognacq Jay Foundation, the Eranet-Neuron (ALTRUISM), the GenMed Labex, AIMS-2-TRIALS which received support from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777394 and the Inception program (Investissement d'Avenir grant ANR-16-CONV-0005). This project has received funding from the European Union's Horizon 2020 research and innovative program CANDY under grant agreement No 847818. The views expressed here are the responsibility of the author(s) only. The EU Commission takes no responsibility for any use made of the information set out.

## AUTHOR CONTRIBUTIONS

T.R. and T.B. designed the research. T.R. performed all analyses, with the help of F.C., R.J.L.A., Al.M., G.H., C.S.L., E.D., A.P., W.K.C., S.J. and T.B. for the genomic analyses, N.T. and R.T for the analysis of brain imaging, and F.A., An.M., and R.D. for the clinical analyses. S.M. developed the website. T.R. and T.B. wrote the manuscript, with assistance from all other authors. All authors approved the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## EXTENDED DATA FIGURES LEGENDS

**Extended Data Figure 1. Framework of the study.**

**Extended Data Figure 2. Admixture results for UK-Biobank, SSC and SPARK cohorts.** For each cohort, the cross-validation errors are shown for increasing values of clusters, and the resulting fraction of ancestry predicted in each admixture group is shown for four clusters. Fractions are shown for reference population individuals and for individuals of each cohorts. The predicted European group, used for subsequent prediction of European ancestry (individuals with > 80% predicted fraction of European ancestry were considered European, see Methods), is shown in dark blue.

**Extended Data Figure 3. Comparison of LOEUF and missense z-score for SPARK genes**. The suggested LOEUF threshold of 0.35 is indicated, as well as the top decile of the missense z-score distribution for all genes, corresponding to a value of 2.45.

**Extended Data Figure 4. Effect of the relative position in encoded protein and of the average pext in brain tissues of high-confidence private singleton LoF variants on ASD relative risk.** Average and standard deviation of relative risk per gene is shown for HC-S-LoFs (**a**) above increasing relative positions of variants on encoded protein and (**b**) below increasing pext scores. Red horizontal bars correspond to a relative risk of 1, and red vertical dashed bars correspond to HC-S-LoF variants that are truncating more than 10% of the encoded protein, i.e. not in the last 10% of the protein sequence, and/or present in more than 10% of the brain-expressed transcripts. Nominal p-values correspond to two-sided Wilcoxon rank tests comparing relative risks at each threshold of relative position or pext score compared to relative risk when all HC-S-LoFs were considered ("Reference"). (**c**) Comparison of gene-level relative risk between HC-R-LoF variants and HC-S-LoF variants. Only genes for which both ASD and control individuals were identified as carrying LoF

variants are considered. Means and standard deviations are shown. P-value from a two-sided Wilcoxon rank test.

**Extended Data Figure 5. Frequency of individuals carrying rare synonymous variants in SPARK genes in each cohort, stratified by status**. Odds ratios (OR) and p-values from two-sided Fisher exact tests. Error bars correspond to standard errors of the proportions.

**Extended Data Figure 6. Examples of LoF variants mapping to exons of SPARK genes**. For *GIGYF1* and *NLGN2*, the LoF variants identified in ASD individuals are indicated on the top, and those identified in control individuals at the bottom.

**Extended Data Figure 7. Prevalence of HC-R-LoF and HC-S-LoF variants.** The fraction of ASD and control individuals carrying HC-R-LoF (top) or HC-S-LoF variants (bottom) are compared (Extended Data Table 3).

**Extended Data Figure 8. Scatterplot of relative risk and attributable risk for HC-R-LoFs.** Relative risk and attributable risk of HC-R-LoFs for each SPARK gene for which at least one variant was found in an ASD individual (Extended Data Table 3). For both relative and attributable risk values, the 95% confidence interval is shown. Inf.: Infinite value due to no variants identified among non-ASD individuals.

**Extended Data Figure 9. Comparison of relative risk to alternative deleteriousness scores.** (**a**) Scatterplot of the HC-S-LoF relative risk for each SPARK gene compared to their LOEUF score. The suggested LOEUF threshold of 0.35 is indicated, as well as the threshold of relative risk of 1. (**b**) Similar representation as (**a**) for HC-R-LoFs. (**c**) Distribution of relative risk for genes above and below the suggested threshold of 0.35. Means and standard deviations are shown. P-value from two-sided Wilcoxon rank test. (**d**) Similar representation as (**c**) for HC-R-LoFs. (**e**) Correlation of HC-S-LoF and HC-R-LoF relative risk

to alternative deleteriousness scores. For each deleteriousness scoring approach, the correlation was measured with relative risk for all SPARK genes for which at least one ASD individual was identified with a variant in our meta-analysis using Pearson correlation tests (95% confidence intervals are shown). The number of SPARK genes considered is indicated. SPARK genes for which none of the control individuals were identified as carrying LoFs were removed for the analysis, and for SPARK genes for which none of the ASD individuals were identified as carrying LoFs, we artificially set the number of carrying ASD individuals to 1 for statistical analyses so that relative risk measures were not infinite (Methods).

**Extended Data Figure 10. Biological processes associated to high relative risk.** (**a**) Correlation between relative risk of HC-R-LoF variants and gene expression in four brain regions (top to bottom: somatosensory, motor and prefrontal cortex; parietal, auditory, visual and temporal cortex; striatum, hippocampus and amygdala; mediodorsal nucleus of the thalamus and cerebellar cortex) and developmental periods for the studied SPARK genes. Correlations and p-values measured by one-sided Kendall correlation tests of correlation between relative risk and gene expression (nominally significant p-values are shown, **nominal $p<0.01$, *nominal $p<0.05$). (**b**) Distribution of HC-R-LoF relative risk for genes encoding synaptic, chromatin organization or transcription proteins compared to relative risk of genes not encoding such proteins. Means and standard deviations of relative risk are shown. Nominal p-values from one-sided Wilcoxon rank tests. (**c**) Similar representation as (**b**) for HC-S-LoFs. For all the tests performed here, SPARK genes for which none of the non-ASD individuals were identified as carrying LoF variants were removed for the analysis, and for SPARK genes for which none of the ASD individuals were identified as carrying LoF variants, we artificially set the number of carrying ASD individuals to 1 so that relative risk measures were not infinite (Methods).

**Extended Data Figure 11. Sex ratio, SCQ and mental distress profiles of HC-S-LoF and HC-R-LoF carriers and non-carriers.** (**a**) The fraction of male and female individuals is

shown for HC-S-LoF variant carriers (left) and non-carriers (right, among ASD individuals (top) and non-ASD individuals (bottom). Odds ratio and p-values for enrichment in male individuals calculated using two-sided Fisher exact tests. SPARK genes for which none of the control individuals were identified as carrying HC-S-LoFs were removed for the analysis. (**b**) Distribution of mental distress history for HC-S-LoF carriers and non-carriers in UK-Biobank control individuals. (**c**) Similar representation as (**a**) for HC-R-LoF variants. (**d**) Distribution of SCQ scores for HC-R-LoF carriers and non-carriers, stratified by status. Means ans standard deviations are shown. (**e**) Similar representation as (**b**) for HC-R-LoF variants.

**Extended Data Figure 12. Association analyses of HC-R-LoF carriers and non-carriers for socio-economical measures.** (**a**) Fraction of HC-R-LoF carriers and non-carriers in each qualification level and income level, and distribution of Townsend Deprivation Index and fluid intelligence score for HC-R-LoF carriers and non-carriers among UK- Biobank control individuals. Means and standard deviations are shown. (**b**) Association analyses of HC-R-LoF carriers and non-carriers for cognitive (fluid intelligence score) and functioning (income, qualification level, Townsend index) measures in the UK- Biobank cohort. The regression coefficients and 95% confidence intervals are derived from linear regression analyses, except for income and qualifications, which were analyzed with ordinal regression analyses (Methods). The direction of effect is adjusted for Townsend index so that higher material deprivation is indicated with a negative sign. P-values are adjusted for multiple testing using the Benjamini and Hochberg control for false discovery rate. All tests were performed when considering genes over increasing thresholds of relative risk, and the number of carriers is indicated.

**Extended Data Figure 13. Brain volume distributions.** For each brain substructure, the mean volume of HC-S-LoF carriers is compared to the mean volume of non-carriers, and the

corresponding standard deviations are shown. P-values from t-tests of difference in mean and Levene tests of difference in variance are shown.

**Extended Data Figure 14. Structural brain anatomy profiles of individuals from UK-Biobank carrying HC-R-LoF variants.** (**a**) Heatmap of the difference in volume in different brain regions between carriers and non-carriers of HC-R-LoFs among UK-Biobank individuals. Clustering of individuals and genes using a hierarchical clustering method with two clusters. All individuals carrying HC-R-LoFs were considered in this analysis. (**b**) Functional enrichment for each gene cluster is measured using two-sided Fisher exact tests and p-values are adjusted using a stochastic approach. Enrichment is shown only for selected terms. (**c**) Four families including multiple individuals carrying the same HC-R-LoF variant are displayed with the age of the individuals, history of mental distress (MD) and z-scored measures of total brain volume and of 13 sub-structures.

**Extended Data Figure 15. Prevalence of HC-S-LoF and HC-R-LoF variants as a function of ASD-PGS.** The ASD-PGS distribution was divided into percentiles and for a subset of percentiles the fraction of individuals carrying a LoF variant over the total number of individuals is shown within the corresponding subgroup, stratified by status and cohort. Nominal p-values and ORs (with 95% confidence intervals) from two-sided Fisher exact tests at the 66$^{th}$ percentile, corresponding to the 3$^{rd}$ tercile of the ASD-PGS distribution (Fig. 2e). All tests were performed for HC-S-LoFs (top) and for HC-R-LoFs (bottom) considering genes over increasing thresholds of relative risk, and the number of variant carriers is indicated.

## EXTENDED DATA TABLES LEGENDS

**Extended Data Table 1. List of 156 SPARK genes studied.**

**Extended Data Table 2. Prevalence of LoF variants in the cohorts studied.** For each cohort, the total number of individuals, the number of LoF carriers and the odds ratio and p-value of enrichment amongst ASD individuals are shown for HC-R-LoFs and HC-S-LoFs, stratified by status and family relationship. OR and p-values from two-sided Fisher exact tests.

**Extended Data Table 3. Fraction of carriers, relative risk and attributable risk by gene.** For each SPARK gene, the fraction of carriers among cases and among controls is provided, as well as the relative risk, attributable risk and corresponding 95% confidence intervals for HC-R-LoF and HC-S-LoF variants. The TADA, pLI and LOEUF scores are also provided.

**Extended Data Table 4. Brain spatiotemporal gene expression levels.** Mean and standard deviation of the expression of all SPARK genes across four brain regions and eight developmental periods are provided.

**Extended Data Table 5. Functional annotation of SPARK genes.** Absence and presence of each SPARK gene in lists of chosen annotations (SynGO biological Process, Gene Ontology Chromatin Organization and Transcription, DNA-templated) is indicated.

**Extended Data Table 6. Statistical tests for differences in brain volume distributions.** For each brain region, the difference in mean (t-test) and standard deviation (Levene test) were calculated between HC-S-LoF carriers and non-carriers. The tests were also performed when considering genes with relative risks higher than 1 and 2.

**Figure 1**

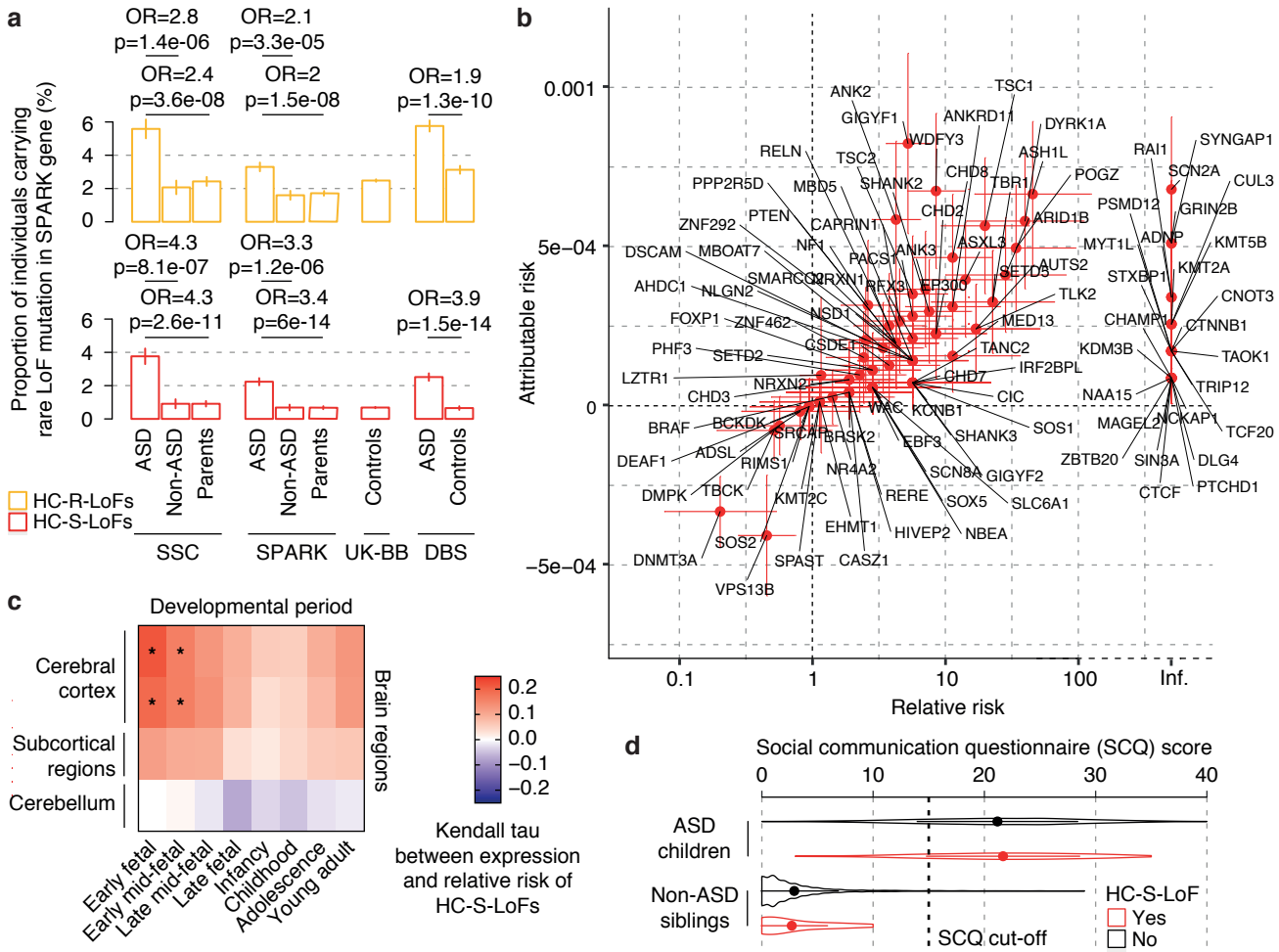
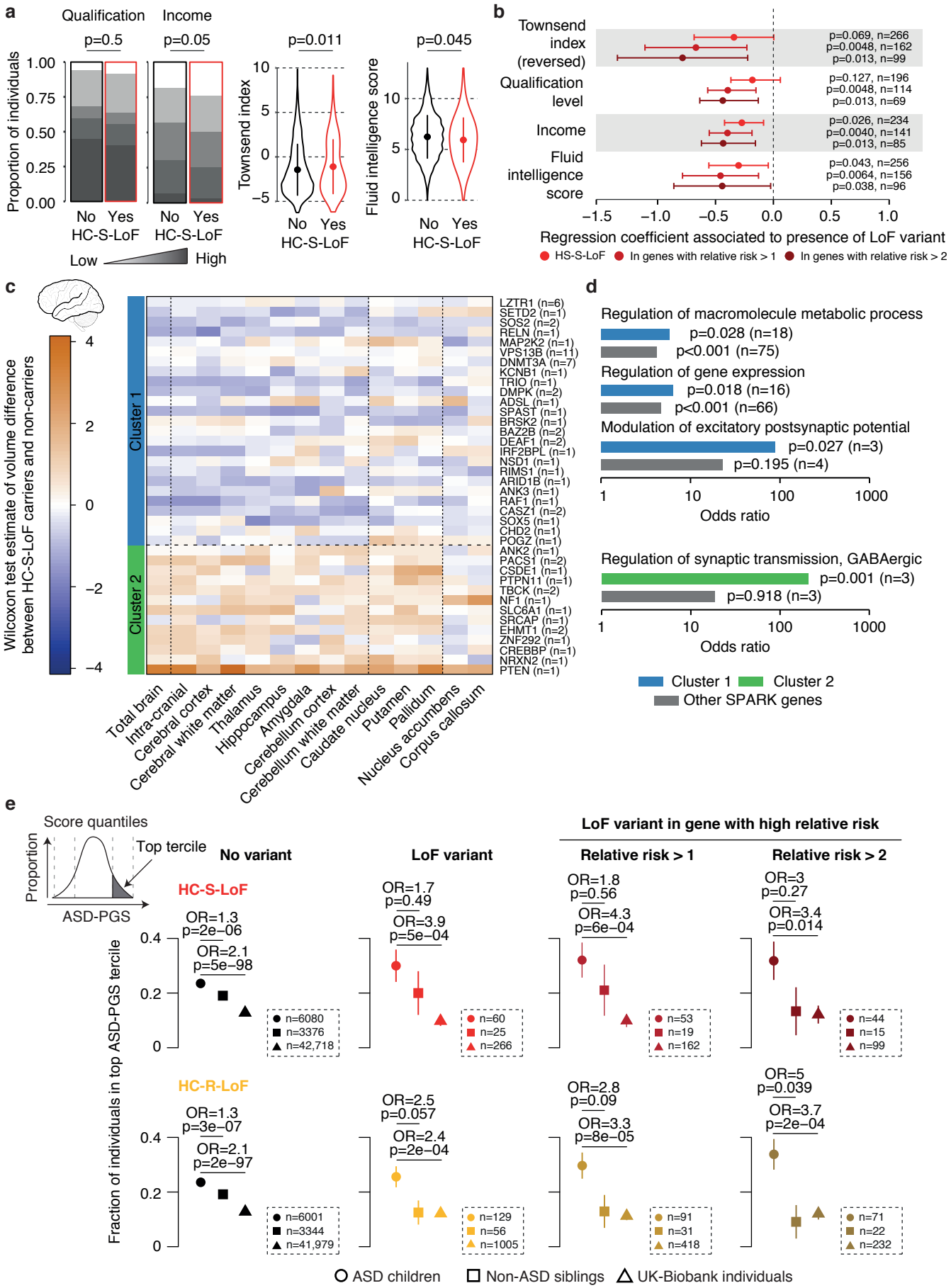

a
OR=2.8
p=1.4e-06
OR=2.1
p=3.3e-05
OR=2.4
p=3.6e-08
OR=2
p=1.5e-08
OR=1.9
p=1.3e-10
OR=4.3
p=8.1e-07
OR=3.3
p=1.2e-06
OR=4.3
p=2.6e-11
OR=3.4
p=6e-14
OR=3.9
p=1.5e-14
Proportion of individuals carrying rare LoF mutation in SPARK gene (%)
ASD
Non-ASD
Parents
Controls
HC-R-LoFs
HC-S-LoFs
SSC
SPARK
UK-BB
DBS
b
Attributable risk
Relative risk
c
Developmental period
Cerebral cortex
Subcortical regions
Cerebellum
Brain regions
Early fetal
Early mid-fetal
Late mid-fetal
Late fetal
Infancy
Childhood
Adolescence
Young adult
Kendall tau between expression and relative risk of HC-S-LoFs
d
Social communication questionnaire (SCQ) score
ASD children
Non-ASD siblings
SCQ cut-off
HC-S-LoF
Yes
No

**Figure 2**


a
Qualification
Income
p=0.5
p=0.05
p=0.011
p=0.045
Proportion of individuals
Townsend index
Fluid intelligence score
No Yes
HC-S-LoF
Low
High
b
Townsend index (reversed)
Qualification level
Income
Fluid intelligence score
p=0.069, n=266
p=0.0048, n=162
p=0.013, n=99
p=0.127, n=196
p=0.0048, n=114
p=0.013, n=69
p=0.026, n=234
p=0.0040, n=141
p=0.013, n=85
p=0.043, n=256
p=0.0064, n=156
p=0.038, n=96
Regression coefficient associated to presence of LoF variant
HS-S-LoF
In genes with relative risk > 1
In genes with relative risk > 2
c
Wilcoxon test estimate of volume difference between HC-S-LoF carriers and non-carriers
Cluster 1
Cluster 2
Total brain
Intra-cranial
Cerebral cortex
Cerebral white matter
Thalamus
Hippocampus
Amygdala
Cerebellum cortex
Cerebellum white matter
Caudate nucleus
Putamen
Pallidum
Nucleus accumbens
Corpus callosum
d
Regulation of macromolecule metabolic process
p=0.028 (n=18)
p<0.001 (n=75)
Regulation of gene expression
p=0.018 (n=16)
p<0.001 (n=66)
Modulation of excitatory postsynaptic potential
p=0.027 (n=3)
p=0.195 (n=4)
Odds ratio
Regulation of synaptic transmission, GABAergic
p=0.001 (n=3)
p=0.918 (n=3)
Cluster 1
Cluster 2
Other SPARK genes
e
Score quantiles
Top tercile
Proportion
ASD-PGS
LoF variant in gene with high relative risk
No variant
LoF variant
Relative risk > 1
Relative risk > 2
HC-S-LoF
OR=1.3
p=2e−06
OR=2.1
p=5e−98
n=6080
n=3376
n=42,718
OR=1.7
p=0.49
OR=3.9
p=5e−04
n=60
n=25
n=266
OR=1.8
p=0.56
OR=4.3
p=6e−04
n=53
n=19
n=162
OR=3
p=0.27
OR=3.4
p=0.014
n=44
n=15
n=99
HC-R-LoF
OR=1.3
p=3e−07
OR=2.1
p=2e−97
n=6001
n=3344
n=41,979
OR=2.5
p=0.057
OR=2.4
p=2e−04
n=129
n=56
n=1005
OR=2.8
p=0.09
OR=3.3
p=8e−05
n=91
n=31
n=418
OR=5
p=0.039
OR=3.7
p=2e−04
n=71
n=22
n=232
Fraction of individuals in top ASD-PGS tercile
ASD children
Non-ASD siblings
UK-Biobank individuals