nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00899-8

Recent ultra-rare inherited variants implicate new autism candidate risk genes

Amy B. Wilfert[1], Tychele N. Turner[1,7], Shwetha C. Murali[1,2], PingHsun Hsieh[1], Arvis Sulovari[1], Tianyun Wang[1], Bradley P. Coe[1], Hui Guo[1,3], Kendra Hoekzema[1], Trygve E. Bakken[4], Lara H. Winterkorn[5], Uday S. Evani[5], Marta Byrska-Bishop[5], Rachel K. Earl[6], Raphael A. Bernier[6], The SPARK Consortium*, Michael C. Zody[5] and Evan E. Eichler[1,2] ✉

Autism is a highly heritable complex disorder in which de novo mutation (DNM) variation contributes significantly to risk. Using whole-genome sequencing data from 3,474 families, we investigate another source of large-effect risk variation, ultra-rare variants. We report and replicate a transmission disequilibrium of private, likely gene-disruptive (LGD) variants in probands but find that 95% of this burden resides outside of known DNM-enriched genes. This variant class more strongly affects multiplex family probands and supports a multi-hit model for autism. Candidate genes with private LGD variants preferentially transmitted to probands converge on the E3 ubiquitin–protein ligase complex, intracellular transport and Erb signaling protein networks. We estimate that these variants are approximately 2.5 generations old and significantly younger than other variants of similar type and frequency in siblings. Overall, private LGD variants are under strong purifying selection and appear to act on a distinct set of genes not yet associated with autism.

Autism spectrum disorder (ASD) is a phenotypically heterogeneous disorder affecting about 1 in 59 children in the United States[1]. Studies to date have primarily focused on high-impact, sporadic variants such as de novo copy number variants (CNVs) and single-nucleotide variants (SNVs). Despite their large effect sizes, DNMs account for approximately 3–25%[2–5] of autism cases. Although this genetic model is highly relevant to simplex ASD, where only one child is affected in a family, it does not explain most cases and is less likely for multiplex families, where more than one child is affected[6]. This has led to the reassessment of various classes of inherited variation and their contribution to autism risk[3,7–12].

It is well established that large, sometimes inherited CNVs underlie a small percentage of sporadic and multiplex autism[3,4]. Preferential transmission of LGD variants have been observed in both simplex[3] and multiplex autism[8] in genes that converge on related functional networks[3,8]. Genetic studies of ASD and developmental delay families have found that affected children are enriched for multiple gene-disruptive variants (CNVs and SNVs)[13–16]. Recent analyses suggest that common inherited risk variants also contribute to ASD pathology[5,17,18].

In this study, we focused on private variants or ultra-rare variants unique to a family. In contrast to other studies[8,11], we did not rely on support from DNMs or DNM rates and excluded genes known to be enriched with DNMs in ASD and neurodevelopmental disorder (NDD) cases to facilitate the discovery of new genes. We expand the number of multiplex and simplex autism families sequenced, taking advantage of the increased sensitivity afforded by whole-genome (WGS) over whole-exome (WES) sequencing data[19] to create a highly sensitive variant callset from WGS data from 3,474 families affected by autism from the Centers for Common Disease Genomics (CCDG) (Table 1 and Supplementary Table 1). We assessed transmission biases in probands and unaffected siblings after controlling for population structure[8,20–23] and replicated these analyses in WES data from 5,879 families from the Simons Foundation Powering Autism Research for Knowledge (SPARK)[24,25]. Our results provide strong support for private LGD variants contributing to autism, particularly multiple hits in different genes. We show that these variants have arisen more recently in autism families (2.5 generations) compared to other classes of variants. Importantly, genes enriched for DNMs contribute little to this burden; rather, we suggest new gene candidates enriched in specific functional pathways.

Results

WGS. We generated WGS data (30-fold sequence coverage) from 2,507 individual DNA samples from 394 families with multiplex autism and 251 families with simplex autism (Table 1, Supplementary Table 1 and Methods). Combined with published WGS data[8,21,22], we created a standardized callset of SNVs and indels from 13,547 genome samples using two callers and made these publicly available (see Data and Code Availability). The set consists of data from 4,364 probands and 2,235 siblings and includes parent–child SNV data from 774 multiplex and 2,700 simplex families. Focusing initially on DNMs, we employed two additional callers and performed 582 random Sanger sequencing validation experiments, combining these with published validation experiments (Supplementary Table 2)[4]. We report an overall validation rate of 99.5% for our DNM callset and estimate a false negative rate of 3.5%. On average, we observed 65.14 DNMs per child and an increase of 1.11 and 0.37 mutations per year of paternal

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [3]Center for Medical Genetic & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, China. [4]Allen Institute for Brain Science, Seattle, WA, USA. [5]New York Genome Center, New York, NY, USA. [6]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. [7]Present address: Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: eee@gs.washington.edu

Table 1 | Summary of WGS and WES

	Individuals				Families				
Cohort	Genomes	Probands	Siblings	Twin pairs	Simplex	Multiplex	Total	Publication	Assay
SSC	8,757	2,299	1,860	1	2,299	0	2,299	An et al.[22]	WGS
SAGE	547	202	5	4	144	26	170	Guo et al.[21]	WGS
TASC	750	250	0	1	250	0	250	Unpublished	WGS
AGRE/iHart phase I[a]	1,736	822	194	69	6	354	360	Ruzzo et al.[8]	WGS
AGRE/iHart phase II	1,757	791	176	0	1	394	395	Unpublished	WGS
Discovery set	13,547	4,364	2,235	75	2,700	774	3,474	-	WGS
SPARK	21,331	6,539	3,034	63	5,278	601	5,879	Unpublished	WES
Combined	34,880	10,905	5,269	164	7,978	1,375	9,353	-	WGS and WES

Summary information of the five cohorts included in our study. WGS samples were sequenced to an average depth of 34×. [a]Ten families had additional members sequenced in AGRE/iHart phase II. These families were still included in the number of simplex/multiplex families for AGRE/iHart phase I; however, each family was identified as simplex or multiplex based on the full family data, which include the additional members from AGRE/iHart Phase II.

and maternal age, respectively (Supplementary Fig. 1). This estimate is lower than reported in Turner et al.[4] because we required three out of four variant callers to agree to increase the specificity of the callset (Methods). As expected, de novo LGD and severe missense mutations (missense with CADD ≥ 30, MIS30) were significantly enriched in probands compared to siblings (LGD odds ratio (OR) = 1.8, $P = 1.43 \times 10^{-23}$; MIS30 OR = 1.25, $P = 1.10 \times 10^{-4}$; Supplementary Table 3). Combining these de novo calls with published DNM callsets (Supplementary Table 4), we identified 100 genes after Benjamini–Hochberg correction (false discovery rate (FDR) < 5%, DNM count > 1) and 45 genes after Bonferroni correction ($P < 5.1 \times 10^{-7}$, DNM count > 1) with an excess of DNMs in autism probands (Methods and Supplementary Table 5). One gene, *MED13*, published as a case report[26], reached Bonferroni significance for an excess of DNMs in autism; four other genes (*CPA6*, *FRA10AC1*, *MPHOSPH10*, *RALGAPB*) reached an FDR < 5% significance. These results replicate the reported excess of de novo LGD mutations in *RALGAPB*[24,27] and identify three genes associated with other neurodevelopmental or neurodegenerative disorders with an excess of de novo missense mutations in ASD. *CPA6* is associated with epilepsy[28], *FRA10AC1* (a gene of unknown function) is associated with Alzheimer's disease[29] and *MPHOSPH10* is associated with early-onset Parkinson's disease[30]. These genes represent excellent candidates for future investigation.

Discovery and properties of private variants. While previous studies focused on the contribution of DNMs or common variants underlying ASD[2–4,6], we focused on the contribution of transmitted variants[3,8,11,12]. Because our previous study showed that transmission disequilibrium signals increased with rarer inherited variants[3], we focused on private inherited variants. We defined these as heterozygous variants observed only once in the parent population and transmitted to at least one child, regardless of potential de novo status in unrelated children within the cohort. Notably, 0.036% of our private variants overlapped with mutations in our DNM callset. Based on our sample size, private variants were ultra-rare in nature and corresponded to an approximate allele frequency $\leq 7 \times 10^{-5}$. We identified 26,606,722 unique private variants (35,871,117 total) in our discovery cohort of 6,599 children and detected no difference in the average number of private variants between probands and unaffected siblings genome-wide (Mann–Whitney *U*-test, autosomes, $P = 0.168$; female X chromosome, $P = 0.328$; male X chromosome, $P = 0.534$). We detected no difference in the number of private autosomal variants transmitted from fathers compared to mothers genome-wide (Wilcoxon signed-rank test, $P = 0.1995$) but detected a difference, as expected, if we considered the female X chromosome (Wilcoxon signed-rank test, $P = 0.0215$; mean paternally transmitted, 98.7; mean maternally transmitted, 102).

Since individuals with similar ancestry have an increased chance of allele sharing compared to individuals from different populations[31], we considered private variants in the context of ancestry. We assigned individuals to one of six super populations (EUR, AFR, AMR, EAS, SAS and OCN) based on maximum likelihood estimations of ancestry using a human diversity panel (Methods and Supplementary Fig. 2). Consistent with previous studies[32,33], children with European ancestry carried the fewest private variants per genome (Fig. 1b,c and Supplementary Tables 6 and 7). This is because most individuals in the discovery cohort (85.6%) were of European ancestry (Supplementary Fig. 2). Private variants among the EUR subgroup are of the lowest frequency, providing the greatest specificity, in principle, to detect pathogenic events[3].

We tested whether filtering against a genetic database (Single Nucleotide Polymorphism Database (dbSNP) v.150) would be sufficient to eliminate this effect and improve our specificity for private events in other populations (Fig. 1c, Supplementary Fig. 3 and Supplementary Table 7). Although dbSNP filtering reduced the average number of private variants per child, the magnitude of the effect varied by population. This treatment reduced the number of private variants by 69.2% among individuals of African ancestry but only 43.6 and 44.9% among individuals of East and South Asian ancestry, respectively. Children of African and East Asian ancestry had, on average, similar variant counts before dbSNP filtering (mean: 11,630 AFR versus 11,653 EAS; Supplementary Table 7). This suggests that the composition of the population genetic database may introduce additional biases because sampling across populations has been nonuniform and allele frequency filtering alone is not sufficient to account for population stratification. These differences highlight the need to evaluate the impact of ancestry and increase underrepresented populations for gene discovery—even in rare variant analyses. We evaluated the impact of population stratification on our results by comparing burden estimates with and without ancestry as a covariate (Supplementary Table 8). We found that the results were comparable and concluded that variation in the number of private variants between populations was generalizable and did not introduce biases into our analyses. Nonetheless, all analyses reported in this study have been replicated in the SPARK cohort and confirmed in the European subset of our discovery cohort (Supplementary Figs. 4–7).

Patterns of private, transmitted variants in protein-coding regions. In this study, we restricted our analyses to autosomal, protein-coding regions of the genome where we expected to have

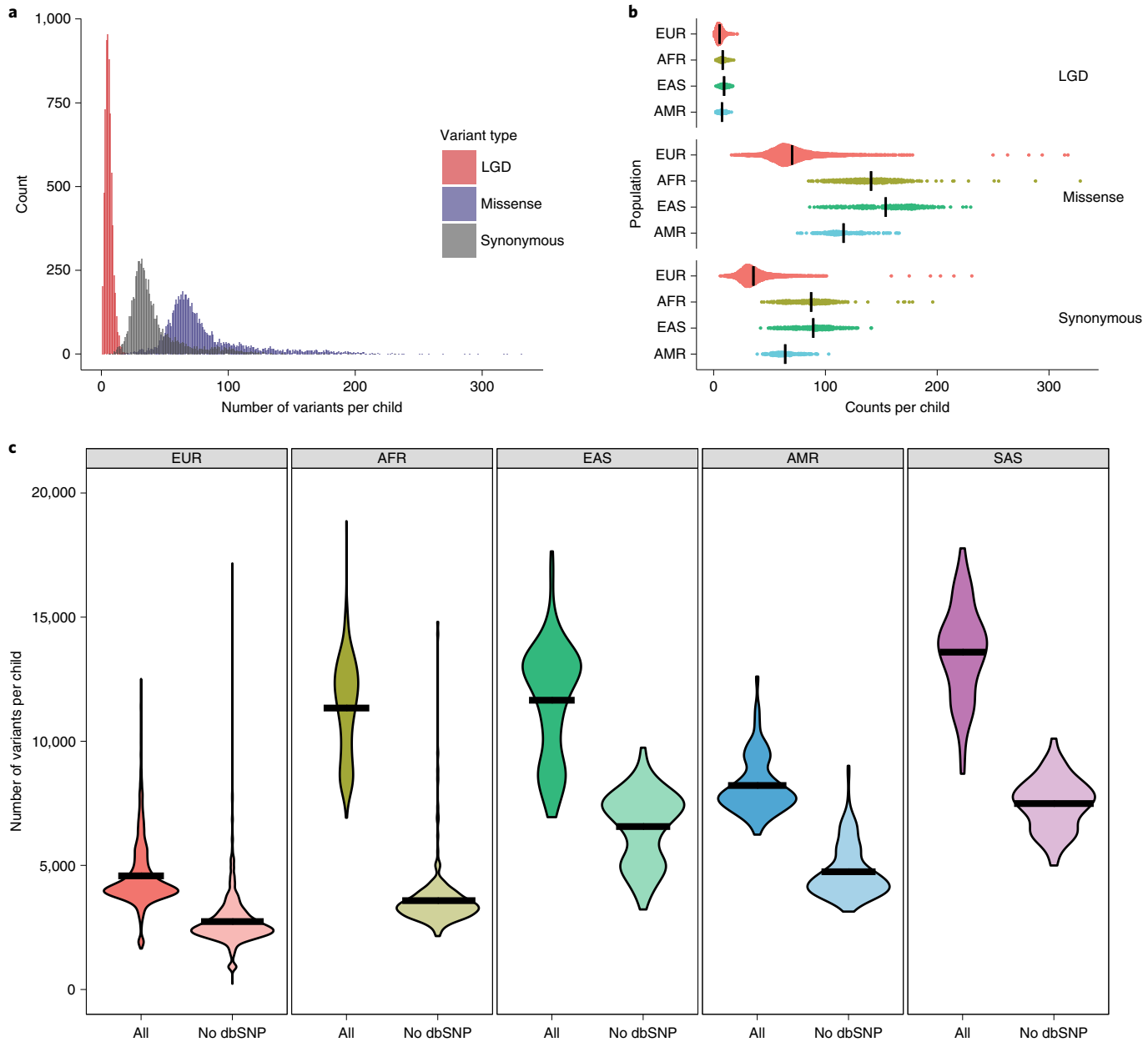

Fig. 1 | Overview of private variants in the discovery cohort. Private variants are defined as variants observed in one and only one parent in the cohort. **a**, Distribution of LGD, missense and synonymous private variants per child (probands and unaffected siblings). **b**, Cumulative number of each variant class by assigned population group (EUR, $n=5{,}685$; AFR, $n=290$; EAS, $n=252$; AMR, $n=193$; SAS, $n=103$), excluding SAS. **c**, Private, transmitted variant counts per child grouped by ancestry before (All) and after (No dbSNP) filtering with dbSNP v.150. Excess of private variants is partially but not fully resolved after excluding sites observed in the dbSNP. We could not assign ancestry to one of these five population groups for 74 of the children in this study. The *y* axis was truncated at 20,000 variants per child; however, both the AFR and EUR populations had a small number of children with variant counts above this threshold (Supplementary Tables 6 and 7). The black lines indicate the average variant count per population in **b** and **c**.

the greatest power to detect enrichment of private, transmitted variants[2,3,8]. Missense variants are the most abundant followed by synonymous and then LGD variants, defined in this study as stop-gain, stop-loss, splice-altering SNVs or frameshift indels (Fig. 1a and Supplementary Table 7). We observed no significant difference between the overall proportion of proband and unaffected sibling carriers for missense, synonymous, or LGD variants and detected no significant enrichment when considering all genes (logistic regression, LGD OR = 1.03, Bonferroni-corrected $P=0.153$; missense OR = 1, Bonferroni-corrected $P=1$; synonymous OR = 1, Bonferroni-corrected $P=1$).

When considering subsets of genes at increasing thresholds of gene constraint using the probability of loss-of-function intolerance (pLI), we replicated[3,8,34,35] the relationship of increasing burden of LGD variants in probands with increasing gene constraint for the discovery, replication, and combined cohorts (Fig. 2a, Supplementary Figs. 4, 8–10, Supplementary Table 9 and Supplementary Note). A similar trend was reported in Satterstrom et al.[11] when considering ultra-rare variants (cohort AC ≤ 5) in a case-control cohort and in approximately 6,305 families. The larger effect size reported in Satterstrom et al. for the case-control cohort is likely due to the presence of DNMs in these samples and the smaller effect size reported

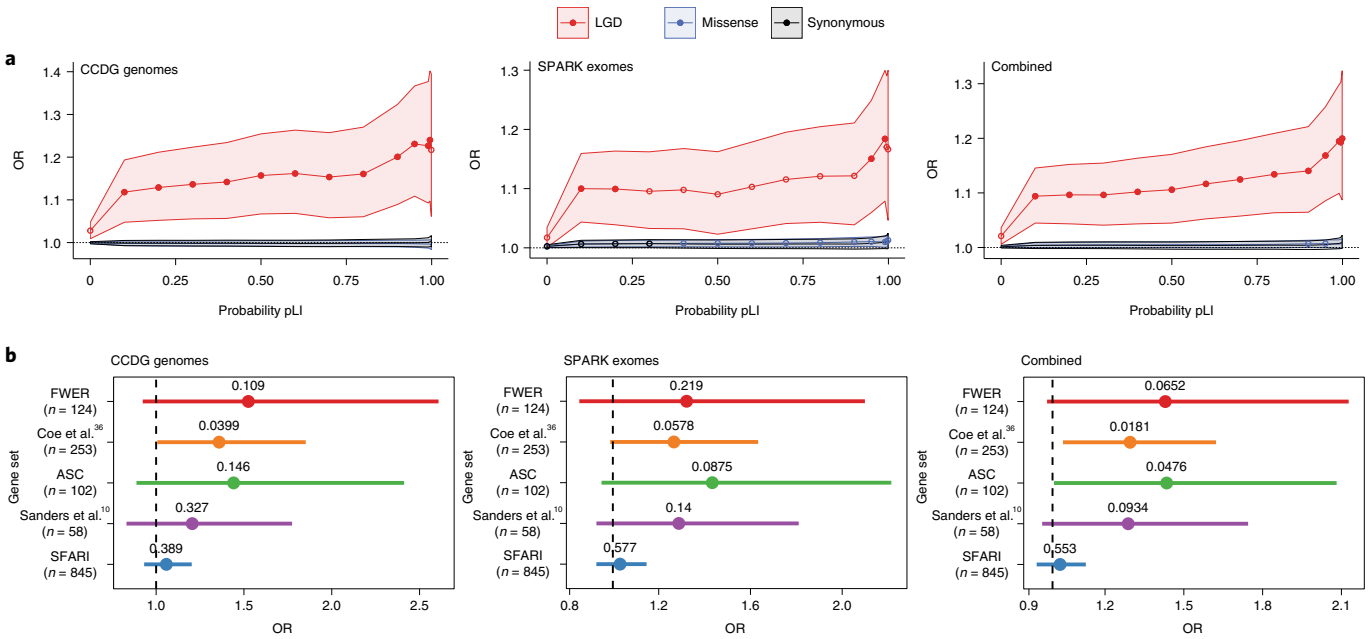

Fig. 2 | Burden of private LGD variants in affected children. a, The burden of private LGD variants in probands compared to siblings was quantified (OR at increasing thresholds of gene constraint (pLI)) in our discovery ($n = 4,201$ affected and 2,191 unaffected children), replication ($n = 6,453$ affected and 3,007 unaffected children) and combined discovery and replication ($n = 10,657$ affected and 5,199 unaffected children) cohorts. The filled circles indicate Bonferroni-corrected $P < 0.05$ (42 tests per cohort); the unfilled circles indicate nominal $P < 0.05$; the shaded areas indicate the 95% confidence intervals (CIs) around the OR estimate. OR and confidence intervals were calculated using logistic regression (Supplementary Table 11). **b**, Enrichment of private, LGD variant transmission to probands for five autism risk gene sets (FWER, Coe et al.[36], ASC, Sanders et al.[10], SFARI). With the exception of SFARI, most gene sets were identified based on an excess of DNMs in parent-child trios (Methods). The OR was based on a comparison of the proportion of carriers between probands and siblings in our discovery ($n = 4,201$ affected and 2,191 unaffected children), replication ($n = 6,453$ affected and 3,007 unaffected children) and combined ($n = 10,657$ affected and 5,199 unaffected children) cohorts using a two-sided Fisher's exact test (Supplementary Table 5). The dashed black line indicates OR = 1, which represents no difference between probands and siblings. Families with monozygotic twins ($n = 75$ in discovery, $n = 63$ in replication and $n = 138$ in combined cohorts) were removed from the analysis. For the combined set, variants were restricted to regions with at least 20× average coverage in the exomes. Reported P values are nominal, points indicate the OR estimate and error bars indicate the 95% CIs around the OR estimate.

for families is likely due to the higher allele frequency threshold used (Supplementary Fig. 11).

We expected this increased burden to be the result of a transmission bias and used a rare variant transmission disequilibrium test to confirm an overtransmission of LGD variants to probands (pLI ≥ 0.99, Bonferroni-corrected $P = 0.0137$ in probands, Bonferroni-corrected $P = 0.52$ in siblings; Supplementary Table 10). Surprisingly, we observed a significant undertransmission of LGD variants in multiplex families with two probands and found no significant increase in allele sharing among affected siblings, suggesting that an LGD variant in one child with autism is not predictive of a second child with autism (Supplementary Fig. 12). We did not observe an increased burden of missense or synonymous variants using pLI gene constraint thresholds. However, we observed an increase in potentially pathogenic, private missense variants in genes with increasing intolerance to missense mutation (Supplementary Fig. 13).

We estimated that the effect size of private, transmitted variants is approximately 8× smaller than the effect size of DNMs (OR = 11.67 for LGD DNMs in genes enriched for DNMs in patients with ASD versus 1.43 for private, transmitted LGD variants in genes with pLI ≥ 0.99). We specifically excluded DNM-enriched genes in cases with ASD as part of this calculation to estimate the effect size excluding well-established genes with an excess of DNMs. In contrast, we compared the burden of private LGD variants among various autism risk gene sets to examine whether the private inherited and DNM signals were exclusive. These included genes shown to be enriched for DNMs in cases with ASD and NDD[10,11,36] and 845 genes from the Simons Foundation Autism Research Initiative (SFARI) (Methods). All gene sets showed a trend toward enrichment of private LGD variants among probands when compared to unaffected siblings but to varying degrees. The Coe et al. gene set[36] showed nominal significance for enrichment in our discovery (Fig. 2b, Supplementary Figs. 5 and 14, Supplementary Table 11 and Supplementary Note; OR = 1.36, nominal $P = 0.040$) and combined cohorts (Fig. 2b, Supplementary Table 11 and Supplementary Note; OR = 1.29, nominal $P = 0.018$). The trends were consistent between replication and discovery cohorts, suggesting that larger sample sizes are required to achieve significance that survives multiple-test correction. In general, DNM-derived gene sets showed greater enrichment than a more general set of autism risk genes (that is, SFARI). DNM-enriched gene sets derived from ASD and NDD studies performed as well (if not better) than those derived strictly from cohorts with autism. Importantly, all trends disappeared if we considered variants at higher allele counts in the parent population (Supplementary Fig. 15), indicating that the signal was strongest for inherited private variants.

Based on the initial sequencing of the SPARK families with autism, Feliciano et al. reported that most of the rare LGD variant transmission bias could not be accounted for by known ASD/NDD genes[24]. We re-evaluated the burden of private, transmitted LGD variants at increasing thresholds of gene constraint, excluding genes enriched for DNMs to quantify this effect. We found that 95.4% of private, transmitted LGD variant burden in probands

remained (Fig. 3a and Supplementary Table 12) at pLI ≥ 0.99 in the discovery cohort. We estimated that private LGD variants in these DNM-enriched genes accounted for 1.45% of ASD risk, whereas private transmitted LGD variants in the remaining genes at pLI ≥ 0.99 accounted for 2.64% of ASD risk (Table 2). Unlike de novo LGD mutations associated with autism, we estimated that most of the attributable risk for private variants awaits discovery and this risk will be identified among genes not already associated with DNM burden. Taken together, these results confirm that DNM-enriched genes confer substantial risk for ASD; however, there is additional burden in the less penetrant set of constrained genes (pLI ≥ 0.99) yet to be discovered.

Simplex versus multiplex and a multi-hit model for ASD. Both our discovery and replication cohorts consist of simplex and multiplex families. Simplex families have been shown[2,37] to be enriched for sporadic or de novo genetic events[20], while multiplex families are more likely to inherit ASD-predisposing variants[38]. We compared the proportion of probands versus siblings carrying at least one private LGD variant at increasing thresholds of gene constraint considering simplex and multiplex families independently (n = 2,700 simplex versus 774 multiplex families; Table 1 and Supplementary Table 1). Despite having 3.5-fold fewer families, multiplex families showed a 25.7% higher burden of private, transmitted LGD variants in probands compared to simplex families, with the greatest effect in less constrained genes (Fig. 3b, Supplementary Figs. 6 and 17 and Supplementary Table 15; multiplex versus simplex OR = 1.37 versus 1.09, permuted P = 0.004 at pLI ≥ 0.1). Among simplex families, significant burden was observed, in contrast, among genes intolerant to mutation (pLI ≥ 0.99).

Previous CNV work and analysis of putative noncoding DNMs[14,19] showed enrichment of multiple deleterious mutations in probands with autism, while other recent studies reported an additive effect between common and rare risk variants[18]. If the signal we observed was relevant to the genetic etiology of autism, we hypothesized that affected children would be more likely to carry multiple private LGD variants, partially explaining why both parents are unaffected or less severely affected in multiplex families. We compared the transmission of two or more private LGD variants in probands and unaffected siblings conditioning on intolerance to mutation. We found that probands were significantly more likely to carry multiple inherited LGD variants in less constrained genes compared to unaffected siblings (Fig. 3c, Supplementary Fig. 18, Supplementary Table 16 and Supplementary Note; OR = 1.29, Bonferroni-corrected P = 0.026 at pLI ≥ 0.1). Under an additive model, which represents independent assortment and random segregation, we would expect the OR for the two-hit model to equal the square of the OR for the one-hit model. This is exactly what we observed and the effect was stronger if we restricted the analysis to individuals of European ancestry (Supplementary Fig. 12), indicating that this signal is not an artifact of population stratification.

New candidate genes and interconnected functional networks. We investigated whether highly constrained genes not enriched for

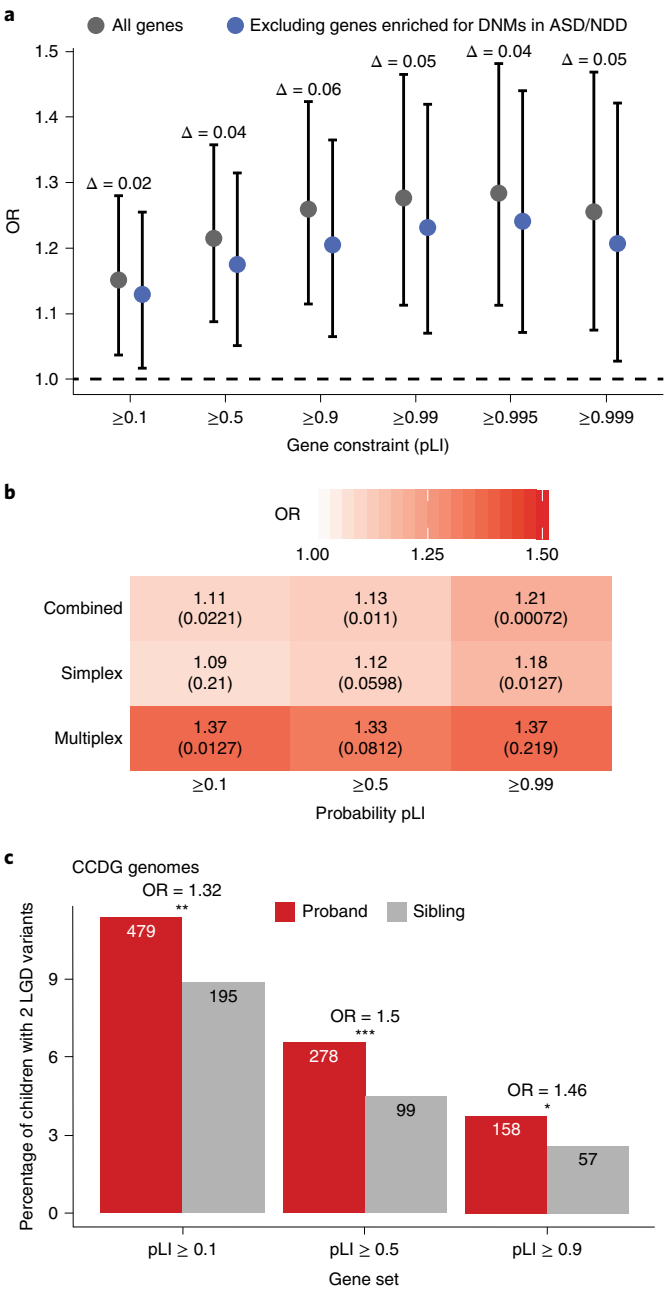

Fig. 3 | Genetic properties of inherited LGD variant burden. a, At least 95.4% of the private, transmitted LGD variant burden resides outside of genes identified with an excess of DNMs in cases with ASD/NDD (321 genes considered and 154 genes with transmissions) based on the analysis of the CCDG autism genomes (n = 4,201 affected and 2,191 unaffected children). We observed 141 DNM-enriched genes with transmissions to probands and 85 genes with transmissions to siblings (Supplementary Table 12). The ORs for five cumulative pLI bins were compared before and after excluding genes with an excess of DNMs in cases with ASD/NDD. The percentage of the remaining burden was calculated as the quotient of the OR for the pLI bin after removing genes enriched for DNMs in cases with ASD/NDD and the OR for all genes in that pLI bin. Families with monozygotic twins (n = 75) were excluded from this analysis. The OR and associated P values were calculated using a two-sided Fisher's exact test. The points indicate the OR estimate and the error bars indicate the 95% CI around the OR estimate. **b**, Multiplex families (n = 1,268 families, 2,691 probands, 533 siblings) showed a higher burden of private, transmitted LGD variants in probands compared to siblings across 3 pLI thresholds compared to simplex families (n = 7,962 families, 7,962 probands, 4,666 siblings). **c**, We observed a significant enrichment of probands carrying two private, transmitted LGD variants when compared to unaffected siblings at various levels of gene constraint (3 cumulative pLI bins considered) based on CCDG genomes sequenced from families with autism (n = 4,201 probands, 2,191 siblings). Families with monozygotic twins (n = 75) were excluded from this analysis. The OR was calculated using a two-sided Fisher's exact test and reported P values were Bonferroni-corrected for nine (**b**) and three (**c**) tests (Supplementary Tables 7 and 8).

DNMs showed enrichment for expression or protein–protein interaction (PPI) networks. Previous studies[8,10,39] typically performed such analyses by integrating candidates with DNM-enriched genes rather than considering them separately. We focused on 163 highly constrained genes (pLI ≥ 0.99; Supplementary Table 17) where private LGD variants are exclusively transmitted to probands and have not been reported in SFARI or as DNM-enriched in three ASD/NDD studies[10,11,36]. Among these genes, a total of 276 LGD variants and 28 genes with independent LGD variants were observed in two or more unrelated families.

Gene ontology (GO) analysis showed that the candidate gene set was highly enriched for encoded phosphoproteins (Supplementary Table 18; KW-0597, 129 out of 163 genes, $q = 1.93 \times 10^{-20}$) and genes were more likely to be interconnected as part of PPI networks (Fig. 4; 102 observed versus 75 expected edges, $P = 0.00164$). A subset of the genes (74 out of 163 genes), including half the genes with events in multiple families, converged on several functional pathways (Fig. 4 and Supplementary Table 18). This included a small network of genes enriched for the E3 ubiquitin ligase pathway by both the GO and Reactome databases, which are involved in proteasome degradation (HSA-98316) and regulation of protein modification by small protein conjugation or removal (GO:1903320). Similarly, a set of more than a dozen genes was associated with internal cellular transport and specifically transport between the Golgi apparatus and endoplasmic reticulum. Other subnetworks were significantly enriched for nucleobase-containing compound metabolic process (GO:006139) and Erb signaling (hsa04012).

This proband candidate gene set was also enriched for cell type-specific expression at the early and mid-fetal cortical stages of human brain development (Supplementary Fig. 19). We observed no enrichment in a set of 83 genes in siblings ascertained using the same criteria (not DNM-enriched, pLI ≥ 0.99, no private LGD variants in probands) (Supplementary Fig. 19). If we focused our expression analyses from brain regions to individual cell types in the adult human cortex, we found that our candidate genes were significantly enriched for expression in both excitatory and inhibitory neurons (Supplementary Fig. 20; excitatory $P = 4.7 \times 10^{-4}$, inhibitory $P = 5.0 \times 10^{-4}$) but not enriched for expression in nonneuronal cell types (Supplementary Fig. 20; $P = 0.24$) compared to control sets. There was no difference between proband and sibling genes ascertained using the same criteria. It should be noted that these pathway enrichments were only observed when compared to the whole genome. If we compared to only genes intolerant to mutation (pLI ≥ 0.99), no pathways were significant.

Private LGD variants in children with autism are evolutionarily younger. Classical population genetics predicts that deleterious variants, such as disease-associated alleles, should be, on average, younger than neutral alleles of the same allele frequency due to purifying selection[40]. Focusing on children of European ancestry, we applied a genome-wide genealogy method developed by Speidel and colleagues[41] that uses the local ancestry (that is, linkage disequilibrium) surrounding a SNP of interest to construct a coalescent tree and estimate the generational age of the allele based on the coalescent branch length. We selected 101 private LGD variants transmitted only to children with autism where none of the 163 candidate genes were previously associated with ASD. We compared them to a random subset of approximately 500 private LGD variants in other genes obtained from both probands and siblings. We estimated the average age of disease-associated LGD variants to be 2.5 generations and found these are significantly younger than other classes of private LGD variants (Fig. 5). We estimated that other proband-associated LGD variants in highly constrained genes (pLI ≥ 0.99) outside the candidate gene set have a median age of 3.7 generations and are significantly older (Mann–Whitney *U*-test, Bonferroni-corrected $P = 0.0133$). Sibling-associated LGD variants in highly constrained genes (pLI ≥ 0.99) were estimated to be almost two generations older (4.3 generations; Mann–Whitney *U*-test, Bonferroni-corrected $P = 0.000255$ candidate versus sibling). As a negative control, we did not observe any difference between the age of private LGD variants in genes outside of the candidate gene set between probands and siblings (Fig. 5; Mann–Whitney *U*-test, $P = 0.139$) or for synonymous or private variants mapping to intergenic regions (Supplementary Figs. 21 and 22). Since alleles in these candidate genes are carried in unaffected parents, we hypothesize that these variants are under weaker selection than deleterious DNMs but under stronger selection than a neutral allele. Specifically, if we assume mutation-selection balance under an additive (for example, two-hit; $h = 0.5$) model[40], we can apply gene-specific mutation rates and allele frequencies within the cohort to estimate the median selection coefficient for the 101 private, transmitted LGD variants in probands of European ancestry. We estimated a rather strong selection coefficient of 0.27 (s.d. = 0.24) for private candidate LGD variants transmitted to only probands with autism in this study.

Table 2 | PAR for de novo and private LGD variants

Variant class	Genes enriched for DNMs in ASD/NDD	Remaining genes, pLI ≥ 0.99
De novo[a]	4.39%	1.45%
Private	1.45%	2.64%

PAR percentages were calculated in our discovery cohort for de novo and private LGD variants in children (Methods). DNM calculations do not include the AGRE study. We defined the DNM-enriched ASD/NDD gene set as the genes reported in Coe et al.[36], Sanders et al.[10] and Satterstrom et al.[11] [a]Does not include the AGRE cohort.

Contribution of known and new ASD-associated variation. Since common variants are implicated in autism risk, we calculated the polygenic risk score (PRS) from our larger sample set and assessed transmission disequilibrium as described recently[17]. We found an even larger difference in the transmission of polygenic risk between probands and unaffected siblings compared to Weiner and colleagues[18], observing the signal in both multiplex and simplex families (Supplementary Fig. 23). We quantified the relative increase in risk for ASD conferred by common variants, DNMs, private SNVs and a set of CNVs previously implicated in autism and NDD. Using a multivariate logistic regression, we estimated the effect size of these four variant categories and computed the population-attributable risk (PAR) (Supplementary Table 19). We restricted this analysis to the Simons Simplex Collection (SSC; $n = 1{,}765$ quads) where we had variant calls across all 4 mutation classes for all samples. We found that children with de novo LGD mutations in DNM-enriched genes were 11.7 times more likely to have autism, accounting for 4.4% of the PAR for ASD. Although CNVs associated with ASD and NDD are, collectively, the rarest events included in this study, children with such an event were 2.7 times more likely to have autism, accounting for 0.9% of the PAR. Carrying one or more private LGD variants in highly constrained genes increases the likelihood of developing autism by 1.4-fold. We estimated that these events account for 3.3% of the PAR, which is comparable to the amount of risk associated with LGD DNMs. Lastly, we found a 1.1-fold increase in the likelihood of developing autism as polygenic risk increases, further supporting a polygenic transmission disequilibrium. While PRS accounts for a large fraction of ASD heritability, we estimated that having a PRS in the top 10% of all children accounts for 1.8% of the PAR for ASD. We note that the contribution of polygenic risk is likely an underestimate since ASD genome-wide association studies to date have been underpowered. Only a small number of robust loci have been identified, so we are likely missing much of the common variant liability for ASD[17]. These 4 categories of risk variants

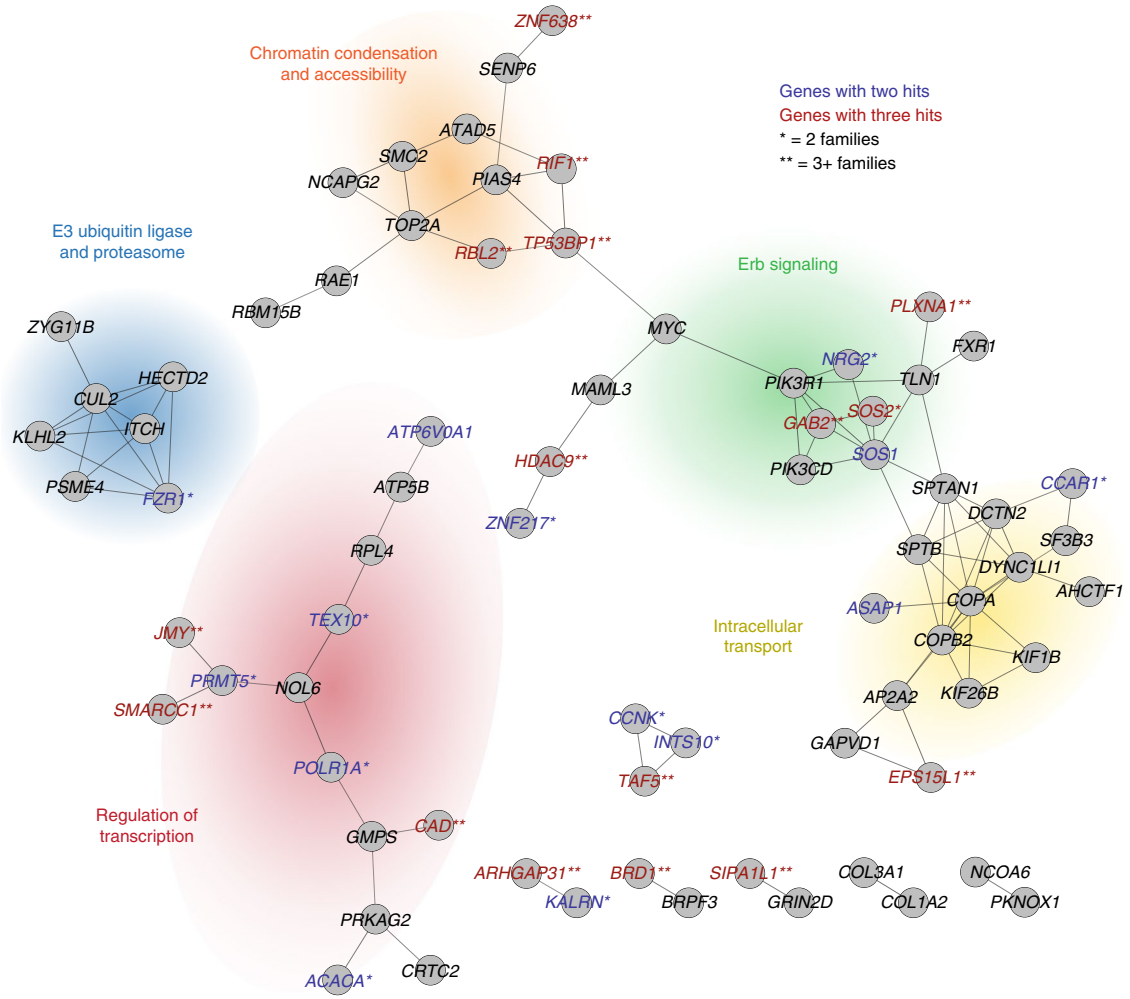

Fig. 4 | PPI network for autism candidate genes. We identified 163 constrained genes (pLI ≥ 0.99) carrying private LGD variants transmitted only to probands with autism based on a combined dataset and not previously identified as a DNM-enriched ASD gene (Supplementary Table 9). STRING network analysis showed a significant excess of PPI ($P = 0.00164$). Gene names are colored if observed in two (blue) or three or more (red) probands and labeled if observed in two independent families (*) or more (**). Families with monozygotic twins ($n = 138$) were removed from the analysis. Analyses were restricted to regions with at least 20× average coverage in the exomes.

only account for 10.4% of the PAR for autism, suggesting that many more risk factors for autism are yet to be discovered.

Discussion

Despite the high heritability of autism, most gene discovery in autism research has been driven by studies of de novo variation[2,8–10,37]. Our analysis shows that ultra-rare transmitted LGD variants are not only significantly enriched in children with autism but contribute to at least 4.5% of autism risk in the human population. This estimate is in line with other studies[3,34] and suggests that this understudied class of variation may confer almost as much risk as de novo SNVs and indels (6–9% of cases using the same PAR estimator)[2,3]. While the burden of private LGD variants in affected children is higher in multiplex families, both simplex and multiplex families showed evidence of biased transmission of private LGD variants. This effect is significant in simplex families only for genes intolerant to mutation, while in multiplex families the effect is larger and significant for genes more tolerant to mutation (Fig. 3b). This may explain why we observed a significant excess of multiple private LGD variants in probands since multiple gene disruptions may be required to reach the diagnostic threshold for ASD.

Some studies focused on identifying risk genes combined de novo and ultra-rare variant risk burden to improve sensitivity, such as the Transmission And De novo Association[39] analysis employed by Ruzzo and colleagues[8]. Because a substantial fraction of DNM-enriched genes have been discovered[8,11,36], we sought to tease apart these effects by excluding known DNM-enriched genes. We estimated that about half of private LGD risk is conferred from genes identified through DNM enrichment studies; excluding known DNM-associated risk genes has a marginal effect on the burden we observed. To enrich for pathogenicity, we identified a set of 163 candidate genes according to gene constraint (pLI ≥ 0.99) and the absence of private LGD variants in unaffected siblings. Although there has been no reported evidence of DNM enrichment in these genes, we found that several of our candidate genes and gene networks identified pathways previously implicated in autism.

For example, we identified three independent private LGD variants in *HDAC9* transmitted exclusively to probands. Pinto et al.[42] identified a transmitted *HDAC9* deletion in a patient with ASD and five additional gene deletions in patients with intellectual disability and schizophrenia, supporting the role of private, transmitted LGD variants in *HDAC9* in ASD pathogenesis. Several other *HDAC* genes have also been implicated in ASD, including *HDAC8* (ref. [43]) and *HDAC4* (ref. [44]) and the chromatin remodeling pathway is known to play a key role in autism[9,37,45]. Another gene in our network, *TOP2A*, is part of the topoisomerase gene family thought

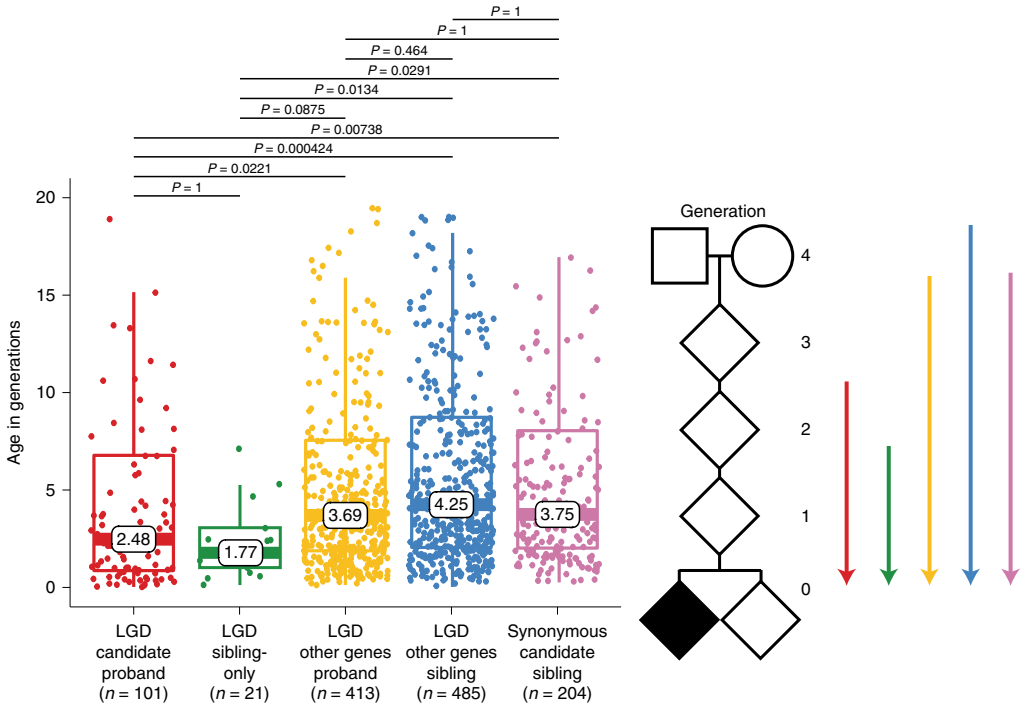

Fig. 5 | Estimate of allele age. The software Relate was used to estimate the coalescent age (in generations) for private LGD (red) and synonymous (blue) variants in 163 candidate genes, private LGD variants (green) in 83 sibling-only genes and approximately 500 sites from all remaining genes for European probands (yellow, $n = 3{,}776$) and siblings (pink, $n = 1{,}909$). *P* values were calculated using a two-sided *t*-test and Bonferroni-corrected for six tests. The plot was truncated at 20 generations. Data points older than this are included in calculating the represented statistics (for example, box plots, median, *P* values) but are not visualized. To view all data points, see Supplementary Fig. 14. Box plot whiskers represent 1.5 times the upper and lower interquartile ranges. The upper and lower hinges correspond to the 25th and 75th centiles; the middle line represents the median. Mean values are noted on the plot.

to be critical in regulating the expression of ASD-related genes[46]. Although this specific topoisomerase has not yet been reported as enriched for DNMs in autism cases, inhibitors of this gene alter the expression of imprinted genes and the topoisomerase acts by resolving transcription-associated supercoiling of long genes, including ASD-related genes critical for synaptic function. Our findings suggest that private inherited variants may identify a subset of genes with variants of smaller effect sizes; however, we and others have shown[36] that more than half of all genes enriched for DNMs have yet to be discovered and none of the inherited genes reach gene-level significance. This is in large part due to sample size and the locus heterogeneity underlying autism. With greater sample sizes, there will likely be more extensive overlap between inherited risk and DNM-risk genes. The case-control study design may be particularly well suited to validate individual candidate genes with an increased burden of private, transmitted variants in autism family studies.

Additionally, we identified a small network of seven genes in the E3 ubiquitin ligase pathway, which has a well-characterized role in autism[47,48]. There are several genes in this pathway enriched for DNMs in children with autism[10,36], indicating that DNMs and private, transmitted LGD variants converge on the same pathway but may be hitting distinct sets of genes. An interesting finding from this study is the discovery of a subnetwork of genes (for example, dyneins, kinesins and coatomer subunits) related to vesicular intracellular transport between the Golgi apparatus and endoplasmic reticulum (Fig. 4). This process is important in the transport of synaptic molecules, such as neurexins and neuroligins, to the cell surface, endocytic cycling of receptors and vesicular cargo transport along microtubules[49–51]. Mutations in related genes in both autosomal recessive and dominant form have been implicated in autism, peripheral neuropathies and NDD. Disruptions in gene function alter synaptic plasticity and morphology of neuronal dendrites and axons. While these associations are exciting, we caution that network and enrichment analyses are often biased toward the most well-studied genes and pathways[52]; thus, more than half the genes that failed to associate with a functional network likely await discovery.

Finally, we report evidence supporting a multi-hit model of autism. We found that private truncating variants in different genes are 50% more likely to occur in probands with autism than siblings, a signal that is consistent with the pathogenicity of this class of variant (Figs. 2a and 3c, Supplementary Table 11 and Supplementary Note). There are other instances of such models reported in ASD ranging from a simple two-hit model[53] to an oligogenic model of disease[4,13,14,19,53]. For example, the 16p12.1 deletion[14,53] is often inherited but requires a secondary CNV to reach the genetic liability threshold for disease. Similarly, carrying 3 or more potentially deleterious DNMs (in the absence of an LGD DNM or large CNV) can be attributed to about 7.3% of cases with autism[4]. A targeted study of seven genes identified a significant overrepresentation of probands with two or more nonsynonymous variants and suggested that multiple moderate impact events in the same pathway are necessary to cause nonsyndromic forms of autism[13]. Efforts focusing on patient recontact, not only for the purpose of re-phenotyping families as diagnostic criteria evolve but also for providing additional counseling as new genetic candidates are identified, will be critical in the task of understanding genotype–phenotype relationships and has already been proposed by others[54]. Understanding the diversity of genetic etiologies underlying autism as well as their corresponding phenotypic outcomes will be critical for providing accurate risk assessments for family planning and genetic counseling.

These findings highlight some key considerations for future ASD studies. Specifically, family composition of the cohort will influence what types of and to what degree different variant classes

contribute to ASD risk. This is important to replicate the findings reported in this study and the findings from other groups[8–10]. Most families with autism characterized by exomes and genomes are simplex in origin; greater effort must be taken to recruit and characterize multiplex families as part of large-scale sequencing efforts. Additionally, these results highlight the weakness of assuming that de novo and rare transmitted variants will impact genes in a similar manner (for example, monogenic and highly penetrant mutations in constrained genes)[35,39,55]. Although we found that de novo and private variants converge on related pathways, our data suggest these two variant classes may act through different genetic mechanisms and modulate distinct sets of genes in ASD pathogenesis. Our allele age estimates are consistent with the action of strong selection operating on these variants. Our analysis suggests that the variants we identified in candidate genes persist for two to three generations before being removed from the gene pool by selection. In contrast, most of the LGD variants associated with de novo-enriched genes are removed from the gene pool almost immediately due to the action of stronger purifying selection.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00899-8.

Received: 19 February 2020; Accepted: 18 June 2021;
Published online: 26 July 2021

References

1. Baio, J. et al. Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. *MMWR Surveill. Summ.* **67**, 1–23 (2018).
2. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
3. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
4. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722.e12 (2017).
5. Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
6. Constantino, J. N. et al. Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. *Mol. Psychiatry* **18**, 137–138 (2013).
7. Ganna, A. et al. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018).
8. Ruzzo, E. K. et al. Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850–866.e26 (2019).
9. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
10. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
11. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584.e23 (2020).
12. Satterstrom, F. K. et al. Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
13. Schaaf, C. P. et al. Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. *Hum. Mol. Genet.* **20**, 3366–3375 (2011).
14. Girirajan, S. et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat. Genet.* **42**, 203–209 (2010).
15. Du, Y. et al. Nonrandom occurrence of multiple de novo coding variants in a proband indicates the existence of an oligogenic model in autism. *Genet. Med.* **22**, 170–180 (2020).
16. Jiang, Y.-H. et al. A mixed epigenetic/genetic model for oligogenic inheritance of autism with a limited role for *UBE3A*. *Am. J. Med. Genet. A* **131**, 1–10 (2004).
17. Grove, J. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
18. 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).
19. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
20. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
21. Guo, H. et al. Genome sequencing identifies multiple deleterious variants in autism patients with more severe phenotypes. *Genet. Med.* **21**, 1611–1620 (2019).
22. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
23. Buxbaum, J. D. et al. The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. *Mol. Autism* **5**, 34 (2014).
24. Feliciano, P. et al. Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genom. Med.* **4**, 19 (2019).
25. SPARK Consortium. SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).
26. Snijders Blok, L. et al. De novo mutations in *MED13*, a component of the Mediator complex, are associated with a novel neurodevelopmental disorder. *Hum. Genet.* **137**, 375–388 (2018).
27. Shah, A. A. et al. Excess of *RALGAPB* de novo variants in neurodevelopmental disorders. *Eur. J. Med. Genet.* **63**, 104041 (2020).
28. Sapio, M. R. et al. Novel carboxypeptidase A6 (*CPA6*) mutations identified in patients with juvenile myoclonic and generalized epilepsy. *PLoS ONE* **10**, e0123180 (2015).
29. Li, Q. S., Parrado, A. R., Samtani, M. N., Narayan, V. A. & Alzheimer's Disease Neuroimaging Initiative. Variations in the *FRA10AC1* fragile site and 15q21 are associated with cerebrospinal fluid $A\beta_{1\text{-}42}$ level. *PLoS ONE* **10**, e0134000 (2015).
30. Siitonen, A. et al. Genetics of early-onset Parkinson's disease in Finland: exome sequencing and genome-wide association study. *Neurobiol. Aging* **53**, 195.e7–195.e10 (2017).
31. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
32. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
33. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
34. Iossifov, I. et al. Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl Acad. Sci. USA* **112**, E5600–E5607 (2015).
35. Epi25 Collaborative. Ultra-rare genetic variation in the epilepsies: a whole-exome sequencing study of 17,606 individuals. *Am. J. Hum. Genet.* **105**, 267–282 (2019).
36. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
37. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
38. Sebat, J. et al. Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
39. He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
40. Maruyama, T. The age of a rare mutant gene in a large population. *Am. J. Hum. Genet.* **26**, 669–673 (1974).
41. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
42. Pinto, D. et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
43. Deardorff, M. A. et al. *HDAC8* mutations in Cornelia de Lange syndrome affect the cohesin acetylation cycle. *Nature* **489**, 313–317 (2012).
44. Williams, S. R. et al. Haploinsufficiency of *HDAC4* causes brachydactyly mental retardation syndrome, with brachydactyly type E, developmental delays, and behavioral problems. *Am. J. Hum. Genet.* **87**, 219–228 (2010).
45. Bernier, R. et al. Disruptive *CHD8* mutations define a subtype of autism early in development. *Cell* **158**, 263–276 (2014).
46. King, I. F. et al. Topoisomerases facilitate transcription of long genes linked to autism. *Nature* **501**, 58–62 (2013).
47. Sanders, S. J. et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).

48. Glessner, J. T. et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
49. Fairless, R. et al. Polarized targeting of neurexins to synapses is regulated by their C-terminal sequences. *J. Neurosci.* **28**, 12969–12981 (2008).
50. Gromova, K. V. et al. Neurobeachin and the kinesin KIF21B are critical for endocytic recycling of NMDA receptors and regulate social behavior. *Cell Rep.* **23**, 2705–2717 (2018).
51. Tomaselli, P. J. et al. A de novo dominant mutation in *KIF1A* associated with axonal neuropathy, spasticity and autism spectrum disorder. *J. Peripher. Nerv. Syst.* **22**, 460–463 (2017).
52. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
53. Girirajan, S. et al. Phenotypic heterogeneity of genomic disorders and rare copy-number variants. *N. Engl. J. Med.* **367**, 1321–1331 (2012).
54. Stessman, H. A., Bernier, R. & Eichler, E. E. A genotype-first approach to defining the subtypes of a complex disease. *Cell* **156**, 872–877 (2014).
55. Epi4K consortium & Epilepsy Phenome/Genome Project. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *Lancet Neurol.* **16**, 135–143 (2017)..

The SPARK Consortium

Xueya Zhou[8,9], Tianyun Wang[1], Pamela Feliciano[10], Jacob Hall[10], Irina Astrovskaya[10], Shwetha C. Murali[1,2], Simon Xu[10], Chang Shu[8,9], Joseph Obiajulu[8,9], Leo Brueggeman[11], Jessica Wright[10], Olena Marchenko[10], Chris Fleisch[10], Timothy S. Chang[12,13], LeeAnne Green Snyder[10], Sarah D. Barns[10], Tychele N. Turner[1,7], Bing Han[10], William Harvey[1], Andrew Nishida[14], Ryan Doan[15,16], Aubrey Soucy[15,16], Brian J. O'Roak[14], Timothy W. Yu[15,16,17], Daniel Geschwind[12,13,18], Jacob Michaelson[11], Natalia Volfovsky[10], Evan E. Eichler[1,2], Yufeng Shen[8] and Wendy K. Chung[9,10]

[8]Department of Systems Biology, Columbia University, New York, NY, USA. [9]Department of Pediatrics, Columbia Medical Center, New York, NY, USA. [10]Simons Foundation, New York, NY, USA. [11]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA, USA. [12]Program in Neurogenetics, David Geffen School of Medicine, Los Angeles, CA, USA. [13]Department of Neurology, University of California Los Angeles, Los Angeles, CA, USA. [14]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [15]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [16]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [17]The Broad Institute, Cambridge, MA, USA. [18]Department of Human Genetics, University of California Los Angeles, Los Angeles, CA, USA.

Methods

Sequencing and quality control of cohorts. Individuals enrolled in the Autism Genetic Resource Exchange (AGRE), SSC, The Autism Simplex Collection (TASC) and Study of Autism Genetics Exploration (SAGE) studies were whole-genome sequenced at the New York Genome Center (NYGC) as part of the CCDG (http://ccdg.rutgers.edu/) (Table 1 and Supplementary Table 1). This study was approved for sequencing by the local institutional review board (IRB) at the NYGC (Biomedical Research Alliance of New York IRB file no. 17-08-26-385). All participants provided written informed consent before participating in the study (SSC: IRB STUDY00001619; SAGE: IRB protocol no. 44219; TASC: STUDY00002514, University of Washington). Sequencing was performed on an Illumina HiSeq X Ten platform using 1 µg of DNA and an Illumina PCR-free library protocol. Post-sequencing, data were processed using the standard pipeline for the CCDG[56] and the GRCh38_full_analysis_set_plus_decoy_hla.fa reference genome. Briefly, raw reads were aligned to the GRCh38 reference genome (Burrows–Wheeler Aligner-MEM v.0.7.15)[57], duplicate reads were marked (Picard v.2.5.0, http://broadinstitute.github.io/picard/), base scores were recalibrated (GATK v.3.8.0)[58] and indels were realigned (GATK). CRAM quality control metrics for the SAGE cohort have been published previously[21]; SSC, TASC and AGRE quality control metrics were determined using Picard WGS metrics, Picard insert size metrics and SAMtools[59] flagstat. The average sequence depths for SSC, TASC and AGRE were 34.99 ± 4.09-fold, 33.89 ± 5.46-fold and 33.03 ± 4.31-fold, respectively. The average insert sizes for SSC, TASC and AGRE were 444.9 ± 17.86 base pairs (bp), 455.1 ± 6.40 bp and 384.8 ± 26.39 bp, respectively.

Individuals enrolled in the SPARK study were whole-exome sequenced at Regeneron (unpublished) (Table 1). Exomes were sequenced to an average coverage of 61.84 ± 14.99-fold. Quality control analysis included HybridizationMetrics (Picard) and SAMtools flagstat.

Variant calling. We called SNVs and indels in families using four different callers: GATK HaplotypeCaller v.3.5.0 (ref. [60]), FreeBayes v.1.1.0 (ref. [61]), Platypus v.0.8.1 (ref. [62]) and Strelka2 v.2.9.2 (Illumina). Additionally, multi-nucleotide variants were called using FreeBayes and Platypus. Post-calling, BCFtools v.1.3.1 (ref. [63]) norm was used to left-align and normalize indels. After variant calling, we partitioned the genome into high-quality regions consisting of unique space and ancient repeats and recent repeat regions, which consisted of repeats <10% diverged from the consensus in RepeatMasker. Variants in high-quality portions of the genome retained for analysis and recent repeat region variants were removed from the study.

Kinship and sample redundancy. All samples from both the discovery and validation cohorts were merged together and kinship coefficients were calculated with KING v.1.4 (ref. [64]). Samples with kinship coefficients that did not match their reported relationship were identified as potential sample swaps or contamination and were either removed or, when possible, their relationships were corrected. Samples with kinship coefficients >0.35 were identified as potential sample duplicates (Supplementary Table 20). We first checked whether potential duplicates were known monozygotic twin pairs or known duplicates within a cohort. (Some individuals had both blood and cell line DNA sequenced for quality control purposes.) We retained one sample from each of the known duplicate pairs, preferentially retaining the sample generated from blood DNA when possible and randomly selecting the retained sample if there was no difference in DNA source. The remaining duplicates, which represented samples sequenced as part of multiple cohorts, were retained for one and only one of the cohorts according to the following prioritization scheme: (1) the sample was sequenced as part of an SSC family; (2) the sample was from a complete family, their DNA was from blood and was WGS; (3) the sample was from a complete family and was WGS; or (4) the sample was from a complete family and contained unaffected siblings. Families with twins were retained for private variant discovery but excluded from all statistical analyses.

Principal component analysis. In addition to our discovery and validation cohorts, we included two reference cohorts, the 1000 Genomes Project (20140818 release)[32] and Simons Genome Diversity Project (SGDP, available at the National Center for Biotechnology Information under BioProject ID PRJNA522307)[65], in our principal component analysis (PCA). Each cohort was cleaned separately (described below), merged together and then cleaned. Reference data from the SGDP and 1000 Genomes Project were prepared for PCA by left-normalizing variants with BCFtools v.1.9, followed by filtering for individual missingness (<10% missing genotypes within an individual), SNP missingness (<50% missing genotypes across an SNP), minor allele frequency (>5%) and linkage disequilibrium pruning with PLINK v.1.90 (ref. [66]). Sites were then converted from hg19 to GRCh38 using the University of California Santa Cruz (UCSC) Genome Browser liftOver. Since the 1000 Genomes Project data were generated with a lower-density SNP array than SPARK, the remaining 1000 Genomes Project sites were the only sites considered in the remaining cohorts. The SSC, SAGE, TASC and AGRE samples were prepared using GATK joint genotype files provided by the NYGC and then iteratively merged together within the respective cohort (most cohorts had to be processed in multiple batches). Each joint genotype file was prepared as described above. The SPARK samples were prepared using the joint genotype generated by Regeneron using Illumina InfiniumCoreExome-24_v1.1 array data. Intensity data files were processed using the Illumina Genome Studio Software v.2.0. Since this dataset was already in PLINK format, it did not undergo additional processing. Before merging all six cohorts together, the 1000 Genomes Project target sites were extracted from each cohort. After merging, the combined autism and reference cohorts were filtered for genotype missingness within the individual and SNP (both < 5%). Finally, the data were input into EIGENSTRAT v.5.0.1 (ref. [67,68]) for PCA. The results of this analysis are summarized in Supplementary Fig. 24.

ADMIXTURE and ancestry assignment. The files we used for PCA input were split by reference (SGDP and 1000 Genomes Project) and discovery cohorts (SSC, SAGE, TASC and AGRE) and filtered for sites present in the reference cohort and individual-level missingness a second time. We ran the software ADMIXTURE v.1.3.0 (ref. [69]) with 10-fold cross-validation on our reference cohort of 1,964 unrelated individuals to determine the optimal value for the *K* parameter. We found that $K = 10$ resulted in the smallest cross-validation error (Supplementary Fig. 2); however, there was little difference in cross-validation error for values of *K* between 8 and 14 and we recognized that a lower value of *K* would result in similar population assignments. We assessed the quality of our inferences for our reference cohort by visualizing the proportion of ancestry from each cluster for a random subset of 15 individuals from each known population (250 individuals in total).

Due to the underlying relationships between individuals in our cohort with autism, we chose to use the allele frequencies learned by ADMIXTURE from our reference cohort to assess the ancestry of our discovery cohort in a supervised manner by using projection with ADMIXTURE. We assigned each individual to the cluster that contributed the largest proportion of ancestry and then grouped clusters into six super populations (EUR, AFR, EAS, SAS, AMR and OCN) according to the membership of known populations from the reference cohort (Supplementary Fig. 2). We were unable to assign ancestry to 1.01% of our discovery cohort due to missing data in the joint genotype files and found that most of our cohort (85.9%) had European ancestry (Supplementary Fig. 2 and Supplementary Table 6).

DNM calls. DNMs in the SSC, SAGE and TASC cohorts were called using a custom pipeline. DNMs were not called in AGRE because DNA for most samples in this cohort was derived from cell lines, which are prone to introducing artifacts in DNM analyses. First, variants that were de novo-based on genotype (the father and mother genotypes were equal to 0/0 and the genotype in the child was 0/1 or 1/1) were retained for further assessment. Second, variants from Platypus with a filter of LowGQX or NoPassedVariantGTs were removed and Strelka2 variants had to have a filter field equal to PASS. Third, variants needed to have the support of at least two of the four callers. Fourth, variants were regenotyped with FreeBayes using default settings and needed to remain as de novo. Fifth, variants in a homopolymer A or T of length 10 or greater were removed. Sixth, we removed all variants in low-complexity regions, recent repeats or centromeres. Finally, we applied the following sample-level filters: the father alternate allele count = 0; mother alternate allele count = 0; child allele balance >0.25; father depth >9; mother depth >9; child depth >9; and either child genotype quality >20 (GATK) or sum of quality of the alternate observations >20 (FreeBayes). For variants on the X chromosome, we considered variants in the pseudoautosomal regions separately (chrX:10,000–2,781,479, chrX:155,701,382–156,030,895) and the X/Y duplicatively transposed region (chrX:89,201,803–93,120,510).

We performed random Sanger validation and combined these data with published validations to look at a total of 3,233 sites in a conditional inference analysis (Supplementary Table 2). The metrics we included in this analysis included: (1) the mer150 mappability, which we calculated on build 38 of the human genome using a workflow originally designed as part of the Encyclopedia of DNA Elements project; (2) the average mapping quality of the read ± 100 bp around the variant in the child; (3) the average mismatch in the reads ± 25 bp around the variant in the child; and (4) the callers that supported the event as de novo. Based on this analysis, the final dataset for de novo SNVs and indels were sites that either had the support of all four callers or were supported by three callers and had an average mapping quality >57 for the reads in the 100-bp region around the variant. For the multi-nucleotide variants, we also inspected all sites using SAMtools tview; the sites had to have visual inspection support of de novo status and an average mapping quality >57 for the reads ± 100 bp around the variant. We estimated our validation rate in this dataset at 99.5% and our false negative rate at 3.5%. In addition, we removed samples that were statistically defined as outliers, in terms of de novo counts, based on the box plot function in R.

Private SNV calls. Each cohort was assessed separately to identify ultra-rare, inherited variants using a custom pipeline. Briefly, SNVs and indels were called using FreeBayes v.1.1.0 and GATK on a per-family basis. Sites were left-aligned, normalized and multiallelic sites were split into separate lines using BCFtools v.1.9. Sites from the two callers were merged using GATK CombineVariants. To ensure a high level of specificity, we counted all alleles in the parent population that were present in the union set of the two callers and passed the following quality control filters: (1) site quality score > 50; and (2) read depth ≥ 10 for genomes

and depth ≥ 20 for exomes. We used slightly different depth filters for the exome and genome data to account for differences in sequencing depth between the two sequencing platforms. All sites that were heterozygous and observed only once in the parent population were designated as candidate private variants (cohort-level parental frequency $\leq 7 \times 10^{-5}$; approximate equivalent ExAC frequency $\leq 2.5 \times 10^{-5}$).

The set of private variants for each cohort consisted of candidate private variants that were present in the intersection set of GATK and FreeBayes and did not violate the rules of Mendelian inheritance. We annotated variants using SnpEff v.4.3t (refs. [70,71]) with gene and transcript information (GRCh38.86), predicted effect of the variant on the transcript, ExAC (release 0.3, non-neuropsych subset) lifted over to GRCh38 using the UCSC liftOver tool and dbSNP v.150. Finally, variants were filtered against recent repeats (see DNM calls for details), low-complexity regions, centromeres and gaps and pseudoautosomal regions (hg38 chrY:10,000–2,781,479, chrY:56,887,902–57,217,415, chrX:10,000–2,781,479, chrX:155,701,382–156,030,895) using BEDTools v.2.24.0 (refs. [72,73]).

The set of private variants from each cohort was compared to all variants observed in the other cohorts. Candidate private variants not observed in any other cohorts were retained for our final set of private variants. For example, the discovery cohort private variants consisted of sites unique to one parent across only the WGS cohorts, whereas the combined set private variants consisted of sites unique to one parent across both the WES and WGS cohorts. When combining the WES and WGS cohorts, we only included regions with an average coverage of 20-fold in the exomes.

Private CNV calls. WES CNVs were called from 21,442 individuals among 5,904 complete families (including 6,582 probands and 3,045 siblings) in the SPARK_WES_1 release using CoNIFER v.0.2.2 (ref. [74]) and XHMM (version statgen-xhmm-3c57d886bc96)[75], as described previously[24]. An independent SNP microarray dataset from 99.2% of the samples (21,271 out of 21,442) was generated using Illumina Infinium Global Screening Array-24v1.0. SNP CNVs were detected using CNVPartition v.3.2.0 (Illumina), PennCNV v.1.0.4 (ref. [76]) and CRLMM v.1.38.0 (ref. [77]) as described previously[24]. We assessed inheritance using both SNP and WES data and filtered putative valid private CNVs based on the inheritance (paternal or maternal), CNV frequency in parents ($n=1$), number of exome probes (>4), percentage of overlap with segmental duplication (<75%) and microarray validation (support by at least one of the CNVPartition, PennCNV and CRLMM approaches). In addition, we required the CNV to only interrupt a single gene to detect 'gene-killing' CNVs. A pLI score was assigned to each gene spanned by a private CNV and CNVs were binned by pLI scores and copy number type (deletions and duplications). A two-sided Fisher's exact test was performed to compare the number of probands and siblings carrying gene-killing private CNVs. Whole-genome structural variant calling was conducted on short-read WGS data from the SSC cohort, consisting of 8,617 samples from 2,276 families, after structural variant calling and quality control, and the SAGE cohort. Calling and merging of structural variants was done as described in Turner et al.[4], with the exception that this study used Delly2 (ref. [78]) instead of VariationHunter[79]. Each genome underwent structural variant calling by six different callers; merging across callers was done in the following order, representing most to least accurate breakpoint callers: WhamG[80], Lumpy[81], Delly2[78], GenomeSTRiP[82], dCGH[83] and CNVnator[84].

De novo burden and enrichment. DNMs were integrated from the CCDG genomes (SSC, SAGE and TASC) in this study, unpublished SPARK exomes (SPARK_WES_1) and three other major published autism exome or genome studies. DNMs, if on hg38, were lifted over to hg19 to enable a merged DNM set. DNMs were restricted to high-coverage (average > 20×) coding regions[85] to combine exome and genome datasets. We also removed mutations that fell within known segmental duplication regions and known recent repeat and low-complexity regions. Sample duplicates, in cohorts like the CCDG genomes and SPARK where the underlying sequencing data were available, were identified using KING v.1.4 to estimate pairwise relatedness between samples. Any samples with a kinship value >0.35 were considered identical and counted only once. Identical samples from the same cohort were also checked for reported monozygotic twin status. Note that samples in SPARK that overlapped with SSC samples were already removed in the final release by the SPARK Consortium. For other published cohorts where the underlying exome data were unavailable, we relied on the published studies to eliminate within-study overlap. For example, we excluded any potential sample overlap across the CCDG genomes in this study with samples in the published literature. This included all SSC samples in the ASC study[15] that overlapped with the CCDG SSC genomes. We also excluded the CCDG TASC genomes ($n=246$) to avoid overlap with the TASC samples in the ASC study and only retained TASC samples in the ASC study ($n=855$) since it has a larger sample size. We further excluded samples with ten or more coding DNMs and removed DNMs seen in five or more different individuals after the above filtering. These measures yielded a total of 15,182 unique ASD trios in the integrated de novo enrichment analysis (Supplementary Table 4). Annotation by variant effect predictor (Ensembl GRCh37 release 94)[86] and combined annotation-dependent depletion score (v.1.3)[87] were applied to ensure uniformity, and the analysis was restricted to the canonical transcript with the most deleterious annotation. A chimpanzee–human divergence model[37] and denovolyzeR[88,89] were used to identify genes with an excess of DNMs. Genes were considered to have an excess of DNMs if both models were significant after multiple-test correction (Benjamini–Hochberg FDR < 5% or Bonferroni correction).

Transmission bias and burden. We partitioned variants from protein-encoding regions of the genome into three classes: (1) likely gene-disrupting, which we defined as any mutation that introduces a stop codon, ablates a stop, changes the frame of the open reading frame, or introduces a change at a predicted splice donor or splice acceptor site; (2) missense, which is any mutation that causes an amino acid change; or (3) synonymous or any mutation that results in no amino acid change. We quantified the number of private, transmitted variants observed in probands and unaffected siblings by gene set and variant type and compared the proportion of carriers using both a Fisher's exact test and logistic regression (one model for each variant type and pLI threshold). For the DNM-enriched gene set analyses, we compared the proportion of carriers between probands and siblings using Fisher's exact test and logistic regression (Supplementary Table 22). Multi-hit analyses and simplex versus multiplex analyses were conducted using a Fisher's exact test to compare the proportion of individuals carrying two or more hits in probands and siblings. We applied Bonferroni and FDR corrections to all P values using the R function p.adjust for each analysis.

Rare variant transmission disequilibrium test. To perform the rare variant transmission disequilibrium test (rvTDT), each child and their parents represented a separate trio. We used the following formula from He et al.[90] to compare the rate of transmitted variants within each set of genes to the expected transmission rate of 0.5. We applied the rvTDT separately to affected and unaffected children. Finally, we calculated an approximate OR based on the fraction of observed transmissions over the expected number of 0.5 transmissions if there was no bias (b = number of children with reference allele transmitted; c = number of children with alternate allele transmitted; thus, the rvTDT: $\chi^2 = (b-c)^2/(b+c)$, d.f. = 1).

PAR. PARs were calculated using the formula published by Cole and MacMahon[91]. Our calculations assumed that siblings are representative of the general population. Even if siblings are subthreshold for ASD, these estimates would serve as a lower bound for PAR, where P_e is the proportion of population (controls) exposed and RR is the the relative risk or ratio of the risk between the exposed and unexposed:

$$\mathrm{PAR}(\%) = \frac{P_e \times (\mathrm{RR} - 1)}{(1 + P_e \times (\mathrm{RR} - 1))}$$

PRS. PRS were calculated using the additive model implemented in PLINK v.1.9 (ref. [66]). Briefly, the genome-wide association study summary statistics from Grove et al.[17] were linkage disequilibrium-pruned; variants with a study $P < 0.01$ were retained for scoring. Each genotype in a sample was weighted with the variant's OR and all of the weighted variants were summed together into a PRS (Supplementary Table 21). To ensure that risk scores were comparable across studies, each cohort was quantile-normalized before combining across cohorts.

Polygenic transmission disequilibrium test. To estimate the burden of common variation in probands and siblings, we performed a polygenic transmission disequilibrium test (pTDT) as described in Weiner et al.[18]. We used the following formulas to calculate the pTDT: n = number of trios/cases; $\mathrm{PRS}_{\mathrm{MP}} = \mathrm{PRS}_{\mathrm{mother}} + \mathrm{PRS}_{\mathrm{father}}/2$; pTDT deviation = $\mathrm{PRS}_{\mathrm{c}} - \mathrm{PRS}_{\mathrm{MP}}/\mathrm{s.d.}(\mathrm{PRS}_{\mathrm{MP}})$; $t_{\mathrm{pTDT}} = \mathrm{mean}(\mathrm{pTDT\ deviation})/\mathrm{s.d.}(\mathrm{pTDT\ deviation})/\sqrt{n}$.

P values were calculated using a two-sided, one-sample t-test with the null hypothesis defined as the mean pTDT deviation equal to zero. All P values were Bonferroni-corrected for 14 tests and 2 conditions (probands and siblings).

Allele age estimation. We estimated the age of a private, transmitted variant at a site of interest using the Relate software v.1.0.16 (ref. [41]). Briefly, Relate reconstructs the local genealogy of the region of interest using a scalable computation, which guarantees the inferred genealogy exactly producing the observed data under the infinite sites model; thus, it can effectively apply to datasets with thousands of individuals while considering recombination. Mutations are then mapped onto the branches of the resulting local tree to estimate mutation age. Each private, transmitted variant is annotated as LGD, synonymous or intergenic, as described above, and then classified into one of the three datasets for candidate gene, affected proband and unaffected sibling depending on the carrier of the private variant. To reduce the computational burden for the inferences for synonymous and intergenic sets, we divided the genome into 100,000-bp windows and randomly selected a locus of interest from up to 550 windows (up to 25 windows per chromosome). We removed sites where the derived allele could not be determined and only included alleles carried by individuals of European descent. For the 163 candidate genes, we included sites within genes where there was a transmitted event within only one family, as well as genes where there were transmitted events of different variants across multiple families. Additionally, we required the genes to carry private LGD variants to have been transmitted only to children with autism and previously associated with ASD. This left us with 101 variants for this gene set.

For a site of interest, we first generated phased haplotypes for the 100,000-bp region surrounding the site of interest for all samples using BEAGLE v.5.1 (ref. [92]) without imputation. For our analysis, we included only one individual from each family, where the family was of European descent, and removed samples that were likely related (see kinship described above). To model recombination wherever applicable, we used the HapMap genetic map. Ancestral and derived states for individual sites were based on the sequences downloaded from Ensembl. To ensure the quality of the genotypes, we masked sites that were within segmental duplications, low-complexity and repeat masker sequences (see private SNV calls described above). The software Relate outputs two age estimates: the lower and upper ages represent the ages of the coalescence events below and above the mutation of interest, respectively. We determined the age of a given variant by taking the average of the two estimates. Note that we recognized that the presence of natural selection at sites where deleterious mutations occur would affect the inference of allele age; thus, the age estimates for deleterious alleles inferred in this study are overestimated and can be deemed as an upper bound for their ages. We compared the distributions of allele age among different datasets using the Mann–Whitney *U*-test; *P* values were all Bonferroni-corrected using p.adjust. Potential caveats, such as phasing errors and cryptic relatedness, might have affected the individual age estimates but were expected to have limited impacts on the observation of differences among the variant sets because the same procedure was applied to individual sets.

Selection coefficient estimation. We applied the classic mutation-selection balance model to estimate the selection coefficients for the 101 private, transmitted variants in the candidate gene set. Our rationale was that because these variants are predominantly recent (on the order of $1\,s^{-1}$)[40], selection must act relatively strongly on these variants; thus, the presence of these variants is primarily due to mutation. Assuming the multi-hit, additive model ($h=0.5$), the individual selection coefficient(s) can be approximated as $mu\,qh^{-1}$, where mu is the gene-specific mutation rate, h is the dominance coefficient and q is the observed allele frequency in the entire population sample.

Gene expression analyses. Cell type-specific expression analyses (CSEAs) were conducted using the CSEA tool[93]. Candidate genes from probands and siblings were uploaded to the available Web server for CSEA across brain regions and development in humans. Gene expression was identified in a published set of transcriptomically defined cell types in the human temporal cortex[94]. Gene sets were tested for enriched expression in three broad cell classes—inhibitory neurons, excitatory neurons and nonneuronal cells—by counting the number of cell types within each class that expressed each gene with average counts per million >1. For each gene set, the number of cell types with expression were calculated and gene sets were visually compared by plotting cumulative distributions. For each cell class, Wilcoxon rank-sum tests were used to identify statistically significant differences in the number of cell types with expression for ASD and control gene sets. *P* values were adjusted for multiple comparisons using Bonferroni correction.

PPI network analysis. We used the STRING database v.11 (ref. [95]) to perform PPI network analyses via Cytoscape v.3.7.2 (refs. [96,97]). We used the multiple protein input option with all default settings except that we required interactions to be limited to high-confidence ones (0.700). Disconnected nodes were hidden from the network output but not from the enrichment analyses. In addition, we also used STRING to calculate network statistics and run functional enrichment analyses from the GO resource, Kyoto Encyclopedia of Genes and Genomes and Reactome pathway database to identify shared functions across the full set of genes and three subnetworks. A subnetwork was identified as any group of genes that contained at least five genes. The genes *KIAA0430* and *ATP5B* are also known as *MARF1* and *ATP5F1B*, respectively.

Statistics. All statistics were calculated using R v.3.5.1 and v.3.5.2.

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

Data availability

The WGS data used in this study are available from the following resources. The AGRE study is available at the Database of Genotypes and Phenotypes (dbGaP) under accession no. phs001766. Access to the AGRE WGS data is subject to approval by Autism Speaks and AGRE. All sequencing and phenotype data for the SSC are available through SFARI and are available to approved researchers at SFARI Base (accession nos. SFARI_SSC_WGS_p, SFARI_SSC_WGS_1 and SFARI_SSC_WGS_2). The genomic and phenotypic data for the SPARK study are available by request from SFARI Base (accession no. SFARI_SPARK_WES_1). Data from the SAGE study are available at the dbGaP under accession no. phs001740.v1.p1. Data from the TASC study are available at dbGaP under accession no. phs001741. Family-level FreeBayes and GATK VCF files for SAGE, SSC and TASC are available under dbGaP accession no. phs001874.v1.p1 and at SFARI Base under accession no. SFARI_SSC_WGS_2a.

Code availability

All software used in this study is publicly available. The code for the ultra-rare transmitted variant pipeline can be found at https://github.com/EichlerLab/ultra_rare_transmitted.git. The code for the figures and analyses are available upon request.

References

56. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
59. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2018).
61. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://arxiv.org/abs/1207.3907 (2012).
62. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
63. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
64. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
65. Hsieh, P. et al. Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes. *Science* **366**, eaax2083 (2019).
66. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
67. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
68. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
69. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
70. Cingolani, P. et al. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3**, 35 (2012).
71. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
72. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1-34 (2014).
73. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
74. Krumm, N. et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22**, 1525–1532 (2012).
75. Fromer, M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
76. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
77. Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R package crlmm for genotyping and copy number estimation. *J. Stat. Softw.* **40**, 1–32 (2011).
78. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
79. Hormozdiari, F. et al. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350–i357 (2010).
80. Kronenberg, Z. N. et al. Wham: identifying structural variants of biological consequence. *PLoS Comput. Biol.* **11**, e1004572 (2015).
81. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
82. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
83. Sudmant, P. H. et al. Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
84. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).

85. Turner, T. N. et al. Sex-based analysis of de novo variants in neurodevelopmental disorders. *Am. J. Hum. Genet.* **105**, 1274–1285 (2019).
86. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
87. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
88. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
89. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo variation in human disease using denovolyzeR. *Curr. Protoc. Hum. Genet.* **87**, 7.25.1–7.25.15 (2015).
90. He, Z. et al. Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. *Am. J. Hum. Genet.* **94**, 33–46 (2014).
91. Cole, P. & MacMahon, B. Attributable risk percent in case-control studies. *Br. J. Prev. Soc. Med.* **25**, 242–244 (1971).
92. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
93. Dougherty, J. D., Schmidt, E. F., Nakajima, M. & Heintz, N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res.* **38**, 4218–4230 (2010).
94. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
95. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
96. Ono, K., Muetze, T., Kolishovski, G., Shannon, P. & Demchak, B. CyREST: turbocharging Cytoscape access for external tools via a RESTful API. *F1000Res.* **4**, 478 (2015).
97. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

Acknowledgements

We thank T. Brown for assistance in editing this manuscript and S. Stray, M. Eng, J. Moore, H. Kortbawi and A. Thornton from the laboratory of Mary-Claire King for the isolation of DNA from whole blood. We thank T. Maniatis and the New York Genome Center for conducting the sequencing and initial quality control. We thank all the families at the participating SSC sites, 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 and E. Wijsman). We thank all the families in SPARK, the SPARK clinical sites and SPARK staff. We appreciate obtaining access to the phenotypic and genetic data on SFARI Base. Approved researchers can obtain the SSC population dataset described in this study (https://www.sfari.org/resource/simons-simplex-collection/) and the SPARK population dataset described in this study (https://www.sfari.org/resource/spark/) by applying at https://base.sfari.org. We gratefully acknowledge the resources provided by the AGRE Consortium and the participating AGRE families. Genomic data for the AGRE cohort was provided by iHART, an initiative led by the Hartwell Foundation and directed by D. Wall and D. Geschwind. This work was supported, in part, by grants from the National Institutes of Health (no. R01 MH101221 to E.E.E.; no. R01 MH100047 to R.A.B.; no. K99 MH117165 to T.N.T.; no. K99 HG011041 to P.H.; and no. UM1 HG008901 to M.C.Z.) and the Simons Foundation (no. SFARI 608045 to E.E.E.). The CCDG is funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute. The Genome Sequencing Program Coordinating Center (no. U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. AGRE is a program of Autism Speaks and is supported in part by grant no. 1U24MH081810 from the National Institute of Mental Health to C. M. Lajonchere. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Author contributions

A.B.W. and E.E.E. designed and conceived the study. L.H.W. and M.C.Z. coordinated the sampling and sequencing for the CCDG cohorts. The SPARK Consortium coordinated the samples and sequencing for the SPARK cohort. A.B.W., T.N.T., S.C.M., A.S., T.W., B.P.C., U.S.E., M.B.-B. and H.G. called the variants and ran the quality control. K.H. performed the Sanger validations. A.B.W., T.N.T., S.C.M., P.H., A.S. and T.W. conducted the analyses and data interpretation. T.E.B. and A.B.W. performed the gene expression analyses. A.B.W., R.A.B. and R.K.E. performed the phenotypic analyses. A.B.W. and E.E.E. wrote the manuscript with input from coauthors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00899-8.

Correspondence and requests for materials should be addressed to E.E.E.

Peer review information *Nature Genetics* thanks Anders Børglum, Thomas Bourgeron and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

nature research

Corresponding author(s): Evan E Eichler

Last updated by author(s): Apr 26, 2021

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	No software was used.
Data analysis	GATK v3.8.0 and v3.5.0, FreeBayes v1.1.0, Strelka2 v2.9.2, Platypus v0.8.1, BWA-mem v0.7.14, King v1.4, ADMIXTURE v1.3.0, Eigenstrat v5.0.1, Relate v1.0.16, R v3.5.2, denovolyzer v0.2.0, SAMtools 1.2-242-g4d56437, Picard v2.5.0, BCFtools v1.3.1, Conifer v0.2.2, XHMM version statgen-xhmm-3c57d886bc96, PennCNV v1.0.4, CRLMM v1.38.0, CNVPartition v3.2.0, Lumpy v0.2.13, Delly2 v0.7.7, GenomeSTRIP v2.0.0, WhamG v1.7.0-296-gb406 dCGH, CNVnator v0.3.3, Plink v1.90, UCSC liftover, Illumina GenomeStudio v2.0, SnpEff v4.3t, BedTools v2.24.0, VEP Ensembl GRCh37 release 94, CADD v1.3, BEAGLE v5.1, STRING v11, CytoScape v3.7.2, SVTyper v0.1.0, https://github.com/EichlerLab/ultra_rare_transmitted.git.

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

The WGS data used in this study is available from the following resources: The AGRE study is available at the Database of Genotypes and Phenotypes (dbGaP) under accession: phs001766. Access to the AGRE WGS data is subject to approval by Autism Speaks and AGRE. All sequencing and phenotype data for the SSC are available through the Simons Foundation for Autism Research Initiative (SFARI) and are available to approved researchers at SFARIbase (http://base.sfari.org, accession IDs: SFARI_SSC_WGS_p, SFARI_SSC_WGS_1, and SFARI_SSC_WGS_2). The genomic and phenotypic data for the SPARK study is available by request from SFARIbase (http://base.sfari.org, accession ID: SFARI_SPARK_WES_1). Data from the SAGE study is available at dbGaP under accession: phs001740.v1.p1. Data from the TASC

study is available at dbGaP under accession: phs001741. Family-level FreeBayes and GATK VCF files for SAGE, SSC, and TASC are available at dbGaP accession phs001874.v1.p1 and also at SFARIbase under accession: SFARI_SSC_WGS_2a. We used data from the following databases: STRINGdb v11, dbSNP v150, 1000 Genomes phase 3, SDGP, Ensembl GRCh37 release 94

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	No sample size calculations were performed. The number of samples used included the maximum number of samples available at the time of the study.
Data exclusions	Samples sequenced as part of multiple studies were excluded from all but one. Poor quality samples as determined by kinship analyses, variant counts per chromosome, and read depth were excluded from analyses.
Replication	All results were replicated once using an independent data set (SPARK) and within the European subset of our discovery cohort. The results of all replication experiments are stated in the manuscript
Randomization	Not relevant - we are not investigating the effect of a treatment/drug
Blinding	Not relevant - we are not investigating the effect of a treatment/drug

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Families of mixed ancestries with unaffected parents and one or more children affected with autism. Genotype information was collected using whole-exome or whole-genome sequencing. Gender and case-control status was collected for all subjects.
Recruitment	Families were recruited by either self-enrollment or physician referral. Families enrolled by self-enrollment may lead to a higher rate of incorrect diagnosis which would lead to our results being an underestimate of the true effect size.
Ethics oversight	This study was approved for sequencing by the local institutional review board (IRB) at NYGC (Biomedical Research Alliance of New York [BRANY] IRB File # 17-08-26-385). All participants provided informed consent prior to participation in the study (SSC: IRB STUDY00001619 [previously, SAGE: IRB protocol #44219, TASC: STUDY00002514 at the University of Washington).

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