nature medicine

Article

https://doi.org/10.1038/s41591-026-04228-6

# Deleterious coding variation associated with autism is shared across ancestries

Received: 20 December 2024

Accepted: 14 January 2026

Published online: 30 March 2026

Check for updates

A list of authors and their affiliations appears at the end of the paper

The past decade has seen remarkable progress in identifying genes that, when impacted by deleterious coding variation, confer high likelihood for autism spectrum disorder (ASD), intellectual disability and other associated developmental disorders. However, most underlying gene discovery efforts have focused on individuals of European ancestry, limiting insights into genetic liability across diverse populations. To help address this, the Genomics of Autism in Latin American Ancestries (GALA) Consortium was formed, presenting here the largest sequencing study of autism in Latin American individuals ($n > 15{,}000$, including 4,717 participants with an ASD diagnosis). We identified 35 genome-wide significant (false discovery rate < 0.05) autism-associated genes, with substantial overlap with findings from European cohorts, and highly constrained genes showing consistent signal across populations. The results provide support for emerging (for example, *MARK2*, *YWHAG*, *PACS1*, *RERE*, *SPEN*, *GSE1*, *GLS*, *TNPO3* and *ANKRD17*) and established autism genes and for the utility of genetic testing approaches for deleterious variants in individuals from diverse backgrounds; the results also demonstrate the ongoing need for more inclusive genetic research and testing. We conclude that the biology of autism is consistent across populations, with no detectable influence of ancestry.

ASD is characterized by deficits in social communication and the presence of restricted interests and/or repetitive behaviors[1]. Although the majority of the genetic liability for autism is attributed to common genetic variation, rare variants, often arising de novo, play a substantial role in individual liability[2,3]. Multiple large-scale studies of rare and common variation associated with autism likelihood are ongoing, and dozens of genes strongly associated with autism have emerged[4,5], primarily coding for proteins involved in gene expression regulation, neuronal communication or the cytoskeleton[6]. These findings have contributed to improved interpretation of genetic tests and represent initial steps in the development of personalized interventions and targeted therapies. Although translation to broad clinical care remains limited, gene-targeted therapeutic strategies for rare genetic disorders associated with autism and other neurodevelopmental disorders (NDDs) have emerged as a very dynamic area of study in both academia and industry[7]. The overwhelming majority of participants in gene discovery studies are of European (EUR) ancestry, even though they comprise only 16% of the global population[8]. This limited window into genetic architecture across ancestries could exacerbate preexisting disparities in diagnostics and service use for autism[9]. Indeed, recent studies have reported high rates of inconclusive results after genetic testing in non-EUR individuals, likely because of uncertainty in interpreting genomic variants[10–12].

We established the GALA Consortium to investigate the impact of genetic and environmental factors on autism across Latin Americans, including participants from all of the Americas, corresponding to the Admixed American (AMR) superpopulation in the 1000 Genomes Project[13]. These AMR individuals comprise the largest recently admixed population in the world and the largest minority in the United States. It is as yet unknown whether the genetic architecture of autism differs across ancestral populations, and the genetic diversity of the AMR group[14,15] makes this question especially relevant.

We present, to our knowledge, the largest sequencing study to date of autism in AMR individuals and compare our results to findings from non-AMR cohorts. We show that a common measure of evolutionary impact on gene-level variation—that is, genomic constraint

e-mail: joseph.buxbaum@mssm.edu

https://doi.org/10.1038/s41591-026-04228-6

scores—differs by ancestry. However, this is not the case for the most constrained genes, which exhibit less population-level variation than expected based on their sequence composition. This is important because most identified autism-associated genes are evolutionarily constrained[4,16], and this applies over diverse populations. Using Bayesian models, we identify 35 genome-wide significant genes associated with autism in Latin American individuals and observe a great degree of overlap with findings in largely EUR cohorts. These results indicate that autism and other NDD genes are shared across ancestries and that existing genetic testing pipelines are effective for the most deleterious variation, especially if information on allele frequency across ancestries is incorporated. We conclude that the biology of autism is consistent across populations and not impacted to any detectable degree by ancestry.

## Results

### Rare variant landscape in Latin Americans diagnosed with ASD

GALA currently encompasses 10 cohorts across the Americas, with data from eight included in this study (Fig. 1 and Methods, 'Description of GALA sites' section). Some GALA samples were contributed to other large-scale whole-exome sequencing (WES) and whole-genome sequencing (WGS) efforts[4,17]; analyses of 1,613 samples (including 707 ASD probands) are reported here for the first time. The GALA analyses reported here include all sequenced samples from GALA cohorts as well as additional genetically inferred AMR samples from the Autism Sequencing Consortium (ASC)[18] and Simons Powering Autism Research (SPARK)[19].

A substantial source of individual autism liability resides in rare deleterious variation in conserved genes[4,6], often de novo or very recent. Hence, to maximize power for discovery, we focus on data collected from trios—that is, an affected proband and both unaffected parents and their typically developing sibling(s), when available. When parental DNA samples could not be collected, we incorporated probands using a case–control framework. After extensive quality control (Extended Data Fig. 1), our analysis included 6,977 individuals: 4,717 ASD cases and the remainder consisting of controls and typically developing siblings (Fig. 1b and Supplementary Table 1). In total, 15,427 individuals were sequenced, including parents from trio-based collections who contributed to de novo variant detection but were not themselves analyzed for variant burden. Specifically, 14,359 individuals were sequenced, with WES ($n$ = 14,152) or WGS ($n$ = 207), as part of family-based analysis: 4,450 AMR ASD individuals, 1,459 siblings and 8,450 parents (Supplementary Table 2). For case–control analysis, 267 ASD AMR samples were matched to 801 non-psychiatric AMR controls from the Mount Sinai BioMe biobank[20,21].

We identified 6,555 rare (that is, allele frequency < 0.1% in our dataset and in the population-specific non-neuro subsets of gnomAD versions 2.1.1 and 3.1.2 (refs. 22,23)) and unique de novo coding sequence variants (5,062 in ASD probands and 1,493 in siblings) (Supplementary Table 3). We identified 36 de novo variants that occurred twice in individuals with ASD: 18 were found in affected siblings, consistent with germline mosaicism, and 18 occurred in unrelated individuals. Additionally, we observed 211 and 15 rare autosomal de novo small genic copy number variants (CNVs)[24] in 2,191 probands and 707 siblings, respectively (Supplementary Tables 4 and 5).

In previous studies, highly constrained genes showed an aggregated signal of variants contributing to autism liability[16], and integrating genomic constraint scores has proven powerful for gene discovery[4]. However, constraint scores are derived from cohorts largely of EUR ancestry. Therefore, we first sought to evaluate the utility of these scores on samples of diverse ancestries.

First, we examined the distribution of de novo variants as a function of a well-established metric of tolerance to loss-of-function variants, the loss-of-function observed/expected upper bound fraction (LOEUF)[22], derived from gnomAD version 2.1.1. Genes with low LOEUF scores are depleted for loss-of-function variation compared to expectation as a result of negative natural selection[22]. Our results demonstrate that rates of de novo variants for both protein truncating (PTV) and deleterious missense (MisB, with a 'missense badness, PolyPhen-2 and constraint' (MPC) score[25] ≥2) variants are elevated in probands compared to typically developing siblings in genes with low LOEUF scores (Fig. 2). Comparing our findings with previously published results[4], we observed that the overall rates of de novo variation in AMR individuals are consistent with those observed in other ancestry groups (Extended Data Fig. 2). Notably, we found a statistically significant enrichment of PTVs in constrained genes among AMR probands. We also observed a trend toward enrichment of missense variants with MPC score ≥ 2 ($P$ = 0.077).

Second, we examined whether LOEUF is well calibrated across ancestral populations. Effective population size differs across Native American, EUR and African populations[26], but current estimates of gene constraint are derived from cohorts that are largely of EUR ancestry. Existing LOEUF scores are modestly over-conservative when applied to AMR samples (Fig. 3, Extended Data Fig. 3 and Karczewski et al.[22]), but, when focusing on the most constrained (lower) deciles, they correlate well with the observed number of PTVs normalized by sample size and gene length (Fig. 3). Because association signal concentrates to these lower deciles (Fig. 2), these observations justify the use of existing LOEUF scores for our study and generally for studies focusing

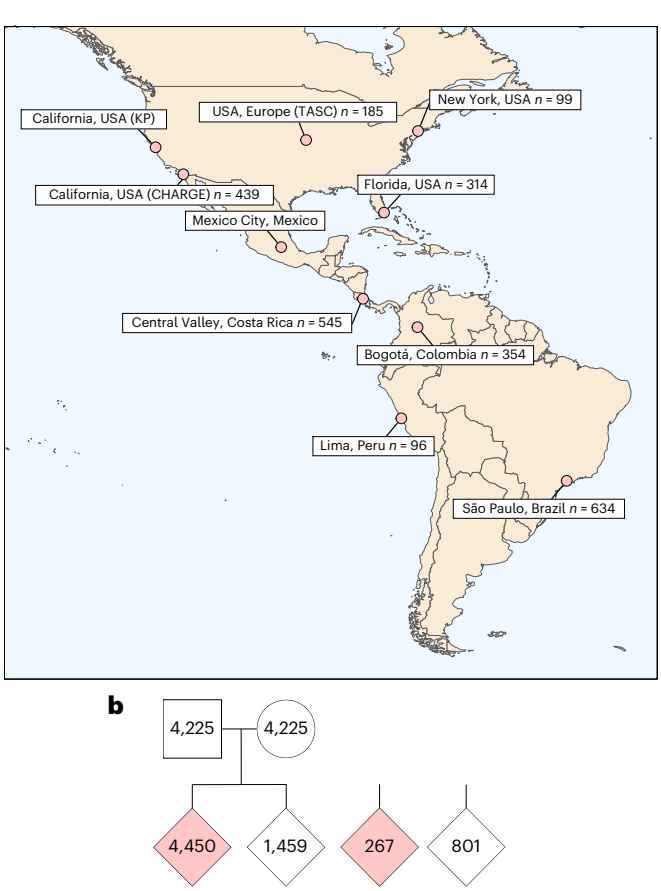



**Fig. 1 | Overview of GALA cohort sites and pedigree structure. a**, Map of GALA collection sites across the Americas. **b**, Pedigree structure of the GALA cohort, comprising 4,717 cases and 10,710 controls. Diamonds represent offspring, and individuals with ASD are shown in pink. Map was generated in R (version 4.3.3) using the ggplot2 and rnaturalearth packages with base map data from Natural Earth (public domain; https://www.naturalearthdata.com/). CHARGE, Childhood Autism Risks from Genetics and Environment; TASC, The Autism Simplex Collection; KP, Kaiser Permanente.

https://doi.org/10.1038/s41591-026-04228-6

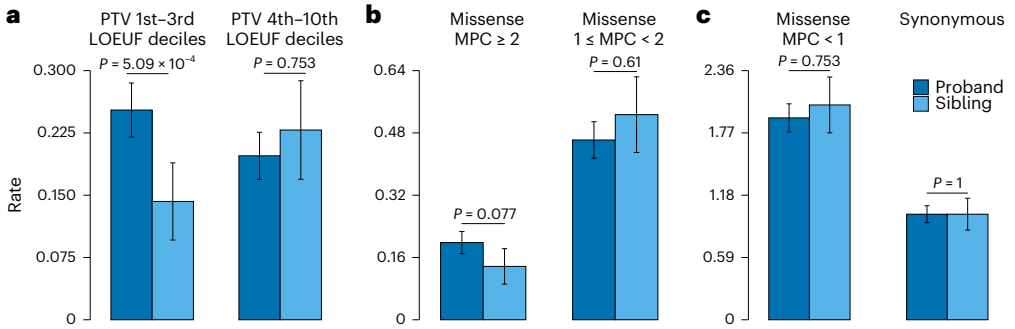



**Fig. 2 | Comparison of rare de novo variant counts per sample between ASD probands and unaffected siblings, normalized to synonymous variant rates.** The average number of rare variants per sample, normalized by the synonymous de novo variant rate, is compared between ASD probands ($n = 4,450$) and unaffected siblings ($n = 1,459$) of AMR ancestry. **a**–**c**, The analysis includes PTVs in highly constrained genes (LOEUF deciles 1–3, 5,363 genes) and less constrained genes (LOEUF deciles 4–10, 12,765 genes) (**a**); missense variants categorized by predicted functional severity (MPC ≥ 2 for high severity, 1 ≤ MPC < 2 for moderate severity) (**b**); and MPC < 1 (for low severity) and synonymous missense variants (**c**). Data are presented as mean values ± 95% CIs. Statistical significance was assessed using two-sided *z*-tests comparing normalized de novo mutation rates between probands and siblings. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg FDR method, and exact adjusted *P* values are shown above the bars.

on highly constrained genes in other ancestries, including admixed African ancestries (Extended Data Fig. 3).

## Autism gene discovery in Latin Americans

For gene discovery, we used TADA (transmission and de novo association), an algorithm that integrates de novo, inherited and case–control variants as well as LOEUF scores and small genic CNVs[4,6,27,28]. Sixteen genes were associated with autism at a false discovery rate (FDR) < 0.01; 35 genes met genome-wide significant association (FDR < 0.05); and 61 genes were associated at FDR < 0.1 (Fig. 4, Table 1 and Supplementary Table 6). To examine the overlap of these findings with those in largely EUR ASD cohorts, we first identified and removed all AMR samples in Fu et al.[4], yielding a non-AMR complementary set ($Fu_{COMP}$) with no overlap with our analyses. Nineteen of the 35 GALA genes with FDR < 0.05 showed significant signal in $Fu_{COMP}$. We next compared the observed numbers of variants in the GALA cohort with the expected number of variants derived from TADA analysis in $Fu_{COMP}$. To do this, we compared results for concordant genes, defined as genes that show FDR < 0.05 in GALA and in $Fu_{COMP}$, and we observed that, overall, the findings are consistent with expectation (Extended Data Table 1). We also compared our gene findings from the GALA cohort with those in a large cohort ascertained for severe developmental disorders[29]: six of the 16 genes that had an FDR < 0.05 in GALA and an FDR < 0.1 in $Fu_{COMP}$ showed an FDR < 0.05 in the developmental disorder cohort (Table 1).

As in previous studies, de novo variation provided a major source of signal for top genes (Extended Data Fig. 4). Similarly, PTVs are a major source of signal, and it was interesting to note that missense variants were also an important source of rare variation association signal (Extended Data Fig. 5). For several of the top genes, the association signal is fully or almost fully derived from missense variants in the GALA cohort, which, for *MTOR*, *YWHAG*, *GRIN1*, *PACS1* and *CACNA1D*, is consistent with previous findings and may suggest a dominant negative or gain-of-function mechanism (Table 1 and Extended Data Table 2). Gene Ontology and Mammalian Phenotype enrichment analyses (Supplementary Tables 7 and 8) highlighted biological processes and phenotypes related to synaptic function, neuronal development and social and repetitive behaviors.

## Implications for clinical genetics

With compelling evidence for overlapping autism gene findings in AMR samples, we next asked about the fraction of findings that are identified as pathogenic or likely pathogenic (P/LP) as per American College of Medical Genetics (ACMG) guidelines[30]. We used VarSome[31]—minimizing the use of proprietary databases and approaches used by commercial testing laboratories—to evaluate (1) genome-wide de novo variation and (2) inherited variation in X-linked genes associated with autism (Supplementary Table 9). This analysis included all de novo variants observed across the genome, not just those meeting the TADA inclusion criteria. Specifically, we included all protein-truncating, missense and synonymous variants, including those in genes lacking mutation rate or LOEUF estimates. For inherited variants, we focused on rare variants in known X-linked genes associated with autism and/or NDDs. We analyzed all GALA and $Fu_{COMP}$ samples, focusing on genes for which there was a reported association with an autism and/or a broader NDD phenotype.

Among the 20,571 de novo variants in our analysis, 926 (4.5%) were classified by VarSome as P/LP when we focused on genes that included autism among the associated phenotypes (Supplementary Table 10). In the AMR cohort, 195 variants (3.8%, 95% confidence interval (CI): 3.27–4.32%) were identified as P/LP (Supplementary Table 11) compared

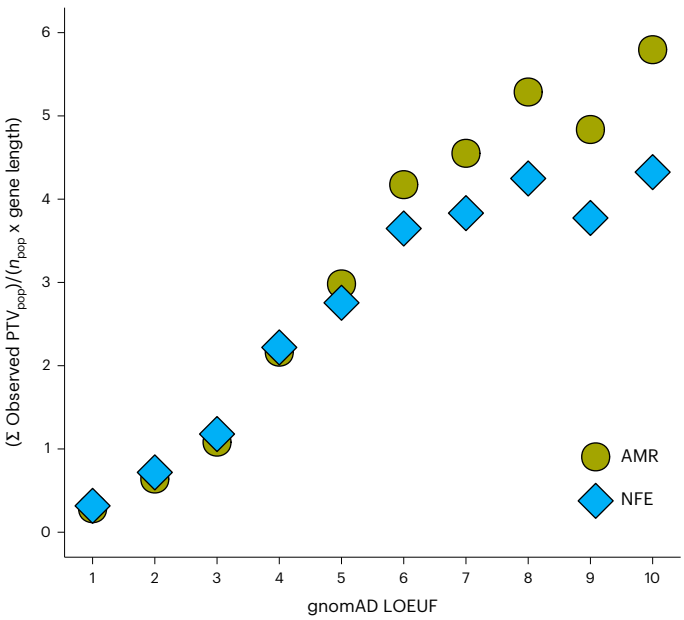



**Fig. 3 | Genic burden of PTVs in EUR versus AMR ancestries as a function of gene constraint.** The sum of observed PTVs is plotted for non-Finnish European (NFE; $n = 56,885$) and AMR ($n = 17,296$) populations in gnomAD version 2.1.1, scaled to population size and total coding sequence length for each gnomAD LOEUF decile. LOEUF deciles reflect gene constraint, with lower deciles indicating more constrained genes.

https://doi.org/10.1038/s41591-026-04228-6

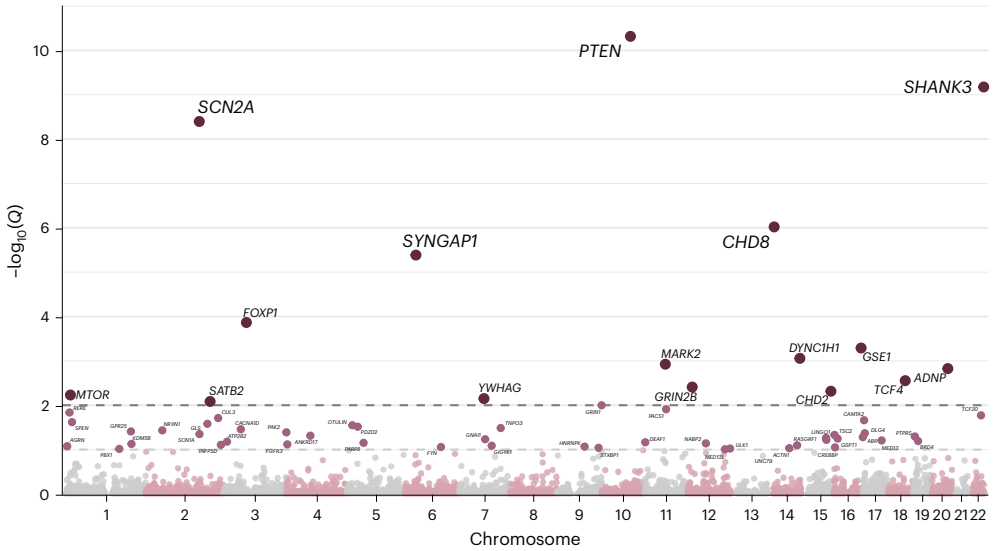



**Fig. 4 | Manhattan plot of autism genes identified in Latin American participants.** The plot displays 16 genes identified with an FDR threshold of <0.01 (dark dashed line) (Table 1), and additional genes identified with an FDR threshold of <0.1 (light dashed line). The top 35 genes, with an FDR <0.05, are shown in Table 1.

to 731 out of 15,386 (4.75%, 95% CI: 4.42–5.10%) in non-AMR samples. In terms of participants with findings, 4.31% (95% CI: 3.75–4.96%) of AMR and 5.53% (95% CI: 5.15–5.94%) of non-AMR probands had at least one P/LP variant identified. Comparisons between EUR and non-EUR participants revealed that EUR individuals had a higher rate of de novo P/LP variants. Specifically, EUR participants had 634 (4.83%, 95% CI: 4.47–5.21%) P/LP variants identified compared to 292 (3.92%, 95% CI: 3.50–4.40%) in non-EUR participants. Overall, EUR participants had a higher rate of P/LP variants identified than non-EUR participants (5.61%, 95% CI: 5.20–6.06% versus 4.54%, 95% CI: 4.05–5.09%).

When broadening our criteria to include other NDD phenotypes, 1,339 de novo variants were deemed to be P/LP (Supplementary Table 12). In AMR, 276 variants were classified as P/LP (5.32%, 95% CI: 4.74–5.98%) versus 1,063 in non-AMR individuals (6.91%, 95% CI: 6.52–7.32%). In terms of participants with de novo findings, 6.07% (95% CI: 5.39–6.82%) of AMR participants and 7.99% (95% CI: 7.53–8.47%) of non-AMR participants had at least one P/LP finding. EUR participants had a notably higher rate of findings (8.22%, 95% CI: 7.72–8.75%) compared to non-EUR participants (6.24%, 95% CI: 5.66–6.87%).

Extending our analysis to include X-linked inherited findings, we observed a further increase in P/LP detection rates. Specifically, 201 de novo or X-linked variants (2.80%, 95% CI: 2.43–3.21%) in AMR samples and 758 variants (3.58%, 95% CI: 3.33–3.84%) in non-AMR samples were classified as P/LP for ASD. When we broadened the terms to include other NDD-related genes, the proportion of P/LP variants rose to 4.10% (95% CI: 3.65–4.58%) in AMR participants and to 5.26% (95% CI: 4.96–5.57%) in non-AMR participants. The rate of participants with at least one P/LP variant increased to 6.47% (95% CI: 5.78–7.24%) in AMR samples and to 8.38% (95% CI: 7.91–8.87%) in non-AMR samples (Supplementary Table 11). EUR participants showed a higher yield of P/LP findings (8.62%, 95% CI: 8.11–9.16%) compared to non-EUR participants (6.63%, 95% CI: 6.04–7.28%).

Qualitatively similar results were obtained when using Neptune[32], which uses databases of previously identified variants to call P/LP variants in a set of 73 ACMG-recommended genes with actionable findings[33] (Extended Data Fig. 6 and Supplementary Table 11). Although greater numbers of rare variants were identified in individuals from diverse ancestries, the proportion of these that could be classified as P/LP was lower. This combination of higher variant detection but reduced classification rate of P/LP variants contributes to a somewhat lower overall yield of P/LP findings per individual in AMR or non-EUR individuals when compared to non-AMR or EUR ancestries, respectively. Considering the VarSome and Neptune results together, the findings provide support for the translatability of rare genetic findings in autism across ancestries in a clinical setting, albeit with opportunities for improvement.

## Discussion

The past decade has seen major advances in deciphering the overall and the genetic architecture of autism but largely from EUR cohorts. It is not yet known whether the genetic architecture of autism differs across ancestral populations, including in admixed populations. Latin American individuals comprise the largest recently admixed population in the world and the largest minority in the United States. Diverse sites with large AMR representation have joined to form GALA, and here we report a first, large-scale multinational analysis of rare variant liability in Latin Americans with ASD, identifying autism-associated genes in this cohort and comparing genetic architecture with that observed in non-AMR ASD.

As in previous studies, we found that signal for genes strongly associated with autism was concentrated in highly conserved genes and largely driven by very rare de novo variation. For the discovery of autism-associated genes impacted by very rare de novo or case–control variation, it is critical to have reliable estimates of expected genic mutation rates, which can be derived from both cross-species comparisons and empirical data from massive, aggregated sequencing resources, such as gnomAD. Although representation of diverse populations is improving, much of the existing sequence data are skewed toward EUR samples. Thus, there is much more to be done regarding genetic variability within underrepresented populations. Our analyses confirm that metrics of gene-level constraint are overly conservative, due to the overreliance on EUR samples that have a lower effective population size. However, we also demonstrate that the key metric LOEUF, when applied to the most conserved genes, is well calibrated across diverse ancestral populations.

Because deleterious variation in highly conserved genes is subject to strong purifying selection, such variation is both very rare and frequently de novo. Allele frequency filtering based on gnomAD or similar datasets is, hence, an important means to infer very rare variation. However, we observe that relying on overall allele frequency allows for the introduction of more common variation into the analyses, hence reducing power and increasing the false-positive rate. We began our

https://doi.org/10.1038/s41591-026-04228-6

**Table 1 | Genome-wide and clinical findings for the top 35 genes**

| Gene | FDR | | | Function | OMIM phenotype | ClinGen disease | Gene2Phenotype mechanism | EAGLE classification | EAGLE score |
|---|---|---|---|---|---|---|---|---|---|
| | GALA | $Fu_{COMP}$ | DD | | | | | | |
| *PTEN* | $4.8\times10^{-11}$ | $6.1\times10^{-14}$ | $<1\times10^{-18}$ | OTH | Macrocephaly/autism syndrome (https://omim.org/entry/605309); Cowden syndrome 1 (https://omim.org/entry/158350) | PTEN hamartoma tumor syndrome | Loss of function | Strong | 63.15 |
| *SHANK3* | $6.7\times10^{-10}$ | $<1\times10^{-18}$ | $<1\times10^{-18}$ | NC | Phelan–McDermid syndrome (https://omim.org/entry/606232) | Phelan–McDermid syndrome | Loss of function | Strong | 74.85 |
| *SCN2A* | $4.0\times10^{-9}$ | $<1\times10^{-18}$ | $<1\times10^{-18}$ | NC | Developmental and epileptic encephalopathy 11 (https://omim.org/entry/613721) | Complex NDD | ID: loss of function; EE: altered gene product structure | Strong | 109.3 |
| *CHD8* | $9.6\times10^{-7}$ | $<1\times10^{-18}$ | $<1\times10^{-18}$ | GER | Intellectual developmental disorder with autism and macrocephaly (https://omim.org/entry/615032) | Complex NDD | Loss of function | Strong | 97.65 |
| *SYNGAP1* | $4.1\times10^{-6}$ | $<1\times10^{-18}$ | $<1\times10^{-18}$ | NC | Intellectual developmental disorder, autosomal dominant 5 (https://omim.org/entry/612621) | Complex NDD | Loss of function | Strong | 40.75 |
| *FOXP1* | 0.0001 | $1.4\times10^{-13}$ | $<1\times10^{-18}$ | GER | Intellectual developmental disorder with language impairment with or without autistic features (https://omim.org/entry/613670) | Intellectual disability, severe speech delay, mild dysmorphism syndrome | Loss of function | Strong | 60.45 |
| *GSE1* | 0.0005 | 0.5834 | 0.275 | GER | | | | | |
| *DYNC1H1* | 0.0009 | $4.8\times10^{-6}$ | $<1\times10^{-18}$ | CYT | Cortical dysplasia, complex, with other brain malformations 13 (https://omim.org/entry/614563) | | Altered gene product structure | Strong | 13.45 |
| *MARK2* | 0.0012 | 0.284 | 0.0002 | NC | Intellectual developmental disorder, autosomal dominant 76 (https://omim.org/entry/621285) | | | Strong | 20.05 |
| *ADNP* | 0.0015 | $<1\times10^{-18}$ | $<1\times10^{-18}$ | GER | Helsmoortel-van der Aa syndrome (https://omim.org/entry/615873) | ADNP-related multiple congenital anomalies–intellectual disability –ASD | Loss of function | Strong | 41.5 |
| *TCF4* | 0.0028 | 0.0096 | $<1\times10^{-18}$ | GER | Pitt–Hopkins syndrome (https://omim.org/entry/610954) | Pitt–Hopkins syndrome | Loss of function | Strong | 13.5 |
| *GRIN2B* | 0.0039 | $9.2\times10^{-11}$ | $<1\times10^{-18}$ | NC | Intellectual developmental disorder, autosomal dominant 6, with or without seizures (https://omim.org/entry/613970); developmental and epileptic encephalopathy 27 (https://omim.org/entry/158350) | Complex NDD | ID: loss of function; EE: altered gene product structure | Strong | 29.65 |
| *CHD2* | 0.0048 | $2.2\times10^{-14}$ | $<1\times10^{-18}$ | GER | Developmental and epileptic encephalopathy 94 (https://omim.org/entry/615369) | Complex NDD | Loss of function | Strong | 25 |
| *MTOR* | 0.0059 | 0.8595 | $2.5\times10^{-10}$ | NC | Smith–Kingsmore syndrome (https://omim.org/entry/616638) | Overgrowth syndrome and/or cerebral malformations due to abnormalities in MTOR pathway genes | Altered gene product structure | | |
| *YWHAG* | 0.0071 | 0.3866 | $4.4\times10^{-6}$ | NC | Developmental and epileptic encephalopathy 56 (https://omim.org/entry/617665) | | Gain of function | | |
| *SATB2* | 0.0082 | 0.0001 | $<1\times10^{-18}$ | GER | Glass syndrome (https://omim.org/entry/612313) | SATB2-associated disorder | Loss of function | Strong | 24.95 |
| *GRIN1* | 0.01 | 0.0071 | $<1\times10^{-18}$ | NC | NDD with or without hyperkinetic movements and seizures, autosomal dominant (https://omim.org/entry/614254) | Complex NDD | Altered gene product structure | | |
| *PACS1* | 0.0123 | 0.104 | $<1\times10^{-18}$ | NC | Schuurs–Hoeijmakers syndrome (https://omim.org/entry/615009) | Schuurs–Hoeijmakers syndrome | Gain of function | Limited | 1.35 |
| *RERE* | 0.0146 | 0.8467 | 0.7127 | GER | NDD with or without anomalies of the brain, eye or heart (https://omim.org/entry/616975) | Complex NDD with or without congenital anomalies | Loss of function | Moderate | 6.5 |
| *TCF20* | 0.0169 | 0.001 | $<1\times10^{-18}$ | GER | Developmental delay with variable intellectual impairment and behavioral abnormalities (https://omim.org/entry/618430) | Developmental delay with variable intellectual impairment and behavioral abnormalities | Loss of function | Strong | 38 |
| *CUL3* | 0.0195 | $2.6\times10^{-5}$ | $5.6\times10^{-9}$ | NC | NDD with or without autism or seizures (https://omim.org/entry/619239) | Complex NDD | Loss of function | Strong | 18.4 |
| *CAMTA2* | 0.0218 | 0.7399 | 0.4596 | GER | | | | Limited | 2.1 |
| *SPEN* | 0.024 | 0.0615 | $1.6\times10^{-11}$ | GER | Radio–Tartaglia syndrome (https://omim.org/entry/619312) | Radio–Tartaglia syndrome | Loss of function | | |

https://doi.org/10.1038/s41591-026-04228-6

**Table 1 (continued) | Genome-wide and clinical findings for the top 35 genes**

| Gene | FDR | | | Function | OMIM phenotype | ClinGen disease | Gene2Phenotype mechanism | EAGLE classification | EAGLE score |
|---|---|---|---|---|---|---|---|---|---|
| | GALA | $Fu_{COMP}$ | DD | | | | | | |
| *GLS* | 0.0262 | 0.7802 | 0.6168 | NC | CASGID syndrome (https://omim.org/entry/618339) | (Infantile cataract, skin abnormalities, glutamate excess and impaired intellectual development: limited) | | | |
| *OTULIN* | 0.0283 | 0.8975 | 0.8426 | NC | (No NDD: autoinflammation, panniculitis and dermatosis syndrome, autosomal dominant; https://omim.org/entry/621030) | | | | |
| *PDZD2* | 0.0305 | 0.6708 | 0.771 | NC | | | | | |
| *TNPO3* | 0.0327 | 0.5381 | 0.0004 | NC | (No NDD: muscular dystrophy, limb-girdle, autosomal dominant 2; https://omim.org/entry/608423) | (No NDD: muscular dystrophy, limb-girdle, autosomal dominant) | | Moderate | 6.35 |
| *CACNA1D* | 0.0348 | 0.6294 | 0.7464 | NC | Primary aldosteronism, seizures and neurologic abnormalities (https://omim.org/entry/615474) | Complex NDD | Gain of function | Strong | 12.7 |
| *NRXN1* | 0.0369 | $4.6 \times 10^{-11}$ | 0.442 | NC | {Schizophrenia, susceptibility to, 17} (https://omim.org/entry/621407) | Complex NDD | Loss of function | Strong | 143.75 |
| *GPR25* | 0.0389 | 0.8914 | 0.8999 | NC | | | | | |
| *PAK2* | 0.0408 | 0.9189 | 0.7868 | CYT | Knobloch syndrome 2 (https://omim.org/entry/618458) | | | | |
| *DLG4* | 0.0428 | 0.0002 | $<1 \times 10^{-18}$ | NC | Intellectual developmental disorder, autosomal dominant 62 (https://omim.org/entry/618793) | Complex NDD | Loss of function | Limited | 2.45 |
| *SCN1A* | 0.0447 | 0.0024 | $<1 \times 10^{-18}$ | NC | Dravet syndrome (https://omim.org/entry/607208); developmental and epileptic encephalopathy 6B, non-Dravet (https://omim.org/entry/619317) | Dravet syndrome; genetic developmental and epileptic encephalopathy | Loss of function | | |
| *TSC2* | 0.0467 | 0.7858 | 0.0217 | NC | Tuberous sclerosis-2 (https://omim.org/entry/613254) | Tuberous sclerosis | Loss of function | | |
| *ANKRD17* | 0.0486 | 0.7463 | 0.3373 | GER | Chopra–Amiel–Gordon syndrome (https://omim.org/entry/619504) | Syndromic complex NDD | Loss of function | | |

For genes with FDR < 0.05 in GALA, we show the FDR in $Fu_{COMP}$[4] and in developmental disorders[29]. In addition, we annotated the genes for function[6]. Only autosomal dominant phenotypes related to NDDs in OMIM or ClinGen are listed. Unless otherwise noted, ClinGen classifications were all 'definitive' or 'strong'. The mutation consequence from Gene2Phenotype and the EAGLE scores[51] from the SFARI Gene database are also included. ASD, autism spectrum disorder; CYT, cytoskeleton; DD, developmental disorder; EE, epileptic encephalopathy; GER, gene expression regulation; ID, intellectual disability; NC, neuronal communication; NDD, neurodevelopmental disorder; OTH, other.

analyses using established best practices for filtering by global allele frequency in the analysis of potentially de novo variants[4,6]. However, we noticed that some variants initially classified as rare in gnomAD (allele frequency < 0.1%) turned out to be more common in particular populations. To address this heterogeneity, we recommend annotating variants with allele frequencies across all subpopulations in the non-neuro releases of gnomAD, as we have done here. Building upon this strategy, we extended the same annotation to our analysis of inherited variation, adopting a more stringent allele frequency threshold of <0.01%, to ensure even more precision in our findings[34,35].

We next used TADA to identify 35 genes associated with autism at an FDR threshold <0.05 in the GALA dataset, 16 with FDR < 0.01 and eight with FDR < 0.001 (Fig. 4 and Table 1). Consistent with previous studies in largely EUR cohorts, gene expression regulation, neuronal communication and cytoplasmic genes are well represented among the autism-associated genes identified in GALA (Table 1 and Supplementary Tables 7 and 8). FDR is well calibrated in TADA[6], and genes identified with TADA in smaller cohorts are consistently replicated at expected levels in larger samples. However, it is still important to evaluate the level of confidence in the genes identified. First, as noted above, we compared results for top genes across GALA and a recent large-scale study (FDR < 0.05 in both AMR samples and non-AMR/$Fu_{COMP}$ studies), and we observed that findings are consistent with expectation (Extended Data Table 1). (Note that, although individual gene-level counts may differ, this variation is expected given the rarity of events; by contrast, when we aggregated data from the top genes, the number of observed variants across genes and variant classes in GALA closely matches the expected total derived from $Fu_{COMP}$.) However, there are multiple genes with evidence in GALA but not in $Fu_{COMP}$. This can be for one of several reasons, including (1) sparseness of de novo events and, hence, overrepresentation/underrepresentation of de novo events in subsamples; (2) differences in ascertainment; and (3) the possibility that some findings are false-positive findings. Although all three could make some contribution, (1) was extensively evaluated previously[4,6], and the analyses suggested that it is likely to be the major contributor to discordance. To further evaluate whether discordant genes may still represent true positives, we first compared GALA findings to results from $Fu_{COMP}$, a non-AMR cohort. Although many top (FDR < 0.05) GALA genes were also supported in $Fu_{COMP}$, a subset of 17 genes showed an FDR > 0.05 in $Fu_{COMP}$, suggesting weaker support. We, therefore, examined their support in a large cohort of individuals with severe developmental disorders[29], and seven of these 17 genes show a clear support. Finally, among the 35 autism-associated genes with an FDR < 0.05, most have a dominant neurodevelopmental morbid association in OMIM, ClinGen and/or Gene2Phenotype (Table 1). The concordance of findings between genome-wide studies (GALA, $Fu_{COMP}$ and developmental disorders) and curated clinical databases indicate that our approach is valid for autism gene discovery in AMR samples and that the FDRs are likely well calibrated.

We next examined emerging and known genes found in the GALA analyses, including contrasting results with those seen in non-AMR samples ($Fu_{COMP}$) and curated databases (Table 1, Fig. 5, Extended Data Table 1 and Extended Data Fig. 7). These genes provide further support for MTOR signaling (for example, *MARK2*, *MTOR*, *TSC2*, *YWHAG* and *GLS*), synaptic and cytoskeletal function (for example, *DYNC1H1*, *PAK2*, *DLG4*, *GRIN1* and *SYNGAP1*) and transcriptional

https://doi.org/10.1038/s41591-026-04228-6

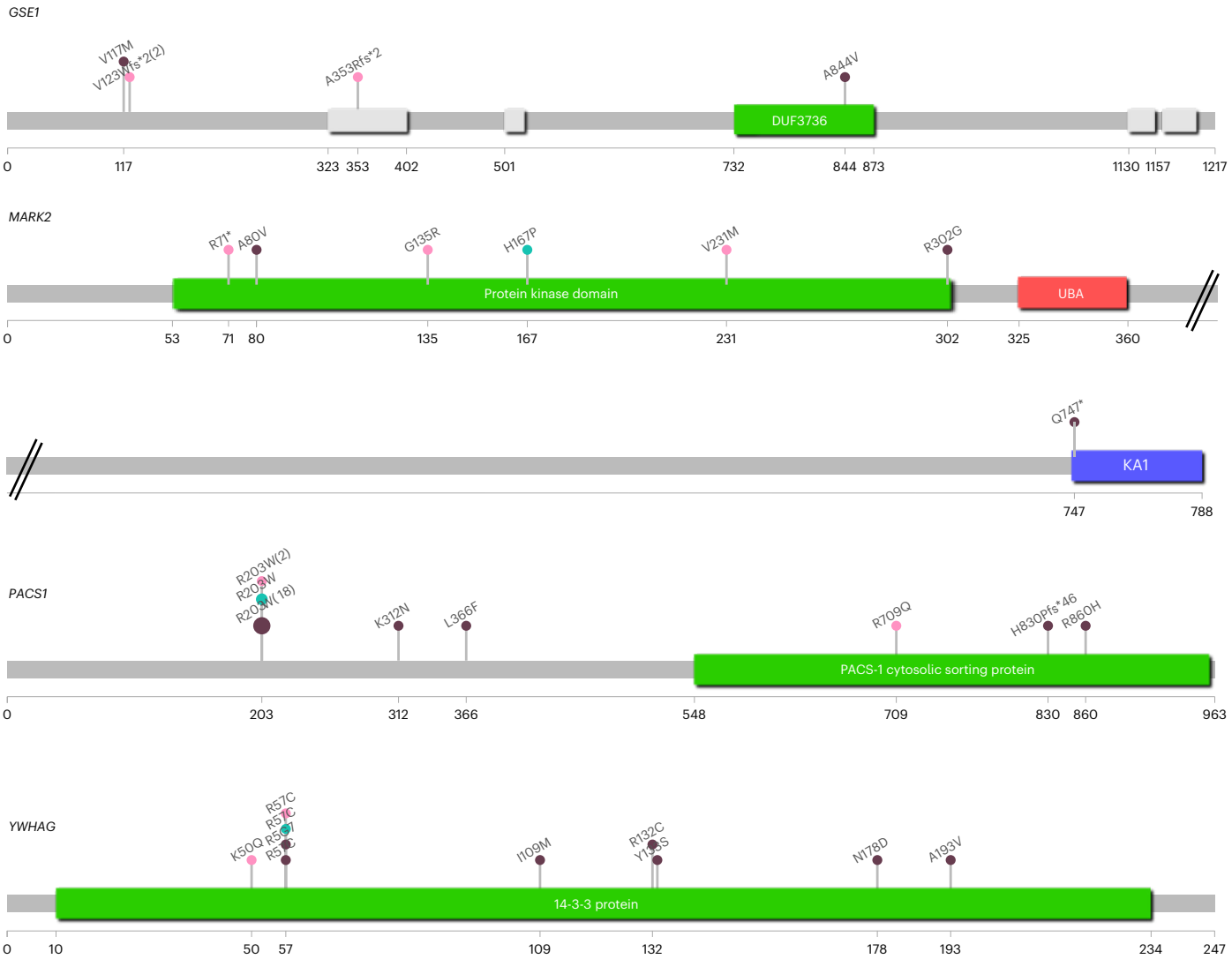



**Fig. 5 | Lollipop diagrams illustrating variants identified in emerging autism-associated genes.** Variants observed in GALA analyses of AMR individuals are marked with pink circles, those found in $Fu_{COMP}$ individuals are marked with green and variants found in DECIPHER are marked with purple. Note that there were two instances of V123Wfs*2 variants in *GSE1* in GALA, two instances of the pathogenic variant R203W in *PACS1* in GALA and 18 in DECIPHER. Figures were generated using the Lollipop software package[52].

regulation (for example, *SPEN*, *RERE* and *GSE1*) in autism. Notably, these pathways are also strongly implicated across intellectual disability and NDDs, underscoring the tremendous overlap in genetic discovery that transcends traditional diagnostic boundaries. Many of these genes are constrained for PTVs and/or missense variants and show support from independent datasets, including de novo events in severe developmental disorder cohorts. A description of top and interesting genes is found in Extended Data Table 2.

Altogether, the results are consistent with the assumption that the same set of highly constrained genes identified in ongoing genome-wide studies is associated with autism, regardless of ancestry. This perspective also receives support from common variant studies in complex traits, where causal effects appear to be highly similar across ancestries[36,37]: Hou et al.[36] analyzed 53,001 African-European admixed individuals and observed that causal effects of common variants (allele frequency > 0.5%) for 38 complex traits are largely similar across local ancestries, in agreement with other studies, including a recent analysis showing that *cis*-genetic effects on gene expression are highly similar between EUR and African individuals[37].

We considered whether the observed similarity in deleterious variant burden between AMR-assigned and EUR-assigned individuals could reflect the influence of EUR admixture within AMR genomes. In principle, local ancestry inference (LAI) would allow mapping of individual variants to ancestral tracts, enabling a more granular test of whether such variants preferentially arise on EUR versus non-EUR backgrounds. However, current LAI methods require dense haplotypic data across the genome, typically from WGS. The sparse and uneven coverage of exome data poses considerable challenges for LAI, and performance has been shown to decline substantially in this context[38,39]. Moreover, because much of our gene discovery relies on de novo rather than inherited variants, the signal is unlikely to be biased by local ancestry tracts, and we also confirmed that variant and gene discovery is clearly driven by the large proportion of individuals with modest overall EUR ancestry (Extended Data Fig. 8). Still, we acknowledge that this is a potential limitation of the study and a valuable direction for future work in cohorts with whole-genome data where LAI can be reliably determined.

Using clinical genetics software platforms, we confirm the overall translatability of clinical genetic approaches when focusing on rare deleterious variation; however, we also reveal differences in the rate of P/LP variants between AMR and non-AMR individuals and between EUR and non-EUR individuals. The causes driving differences in rates of P/LP need to be better understood, as this is a limitation that complicates

the interpretation of our analyses. A recent study focusing on pediatric patients with serious neurologic, cardiac or immunologic conditions reported similar diagnostic yield for genome sequencing in European Americans and Latin Americans (19.8% versus 17.2%); however, yields were lower (11.5%) and inconclusive results were higher in African Americans[11]. In that study, genome sequencing was carried out by commercial diagnostic laboratories, making use of a proprietary pipeline that incorporates variant databases; the degree to which proprietary algorithms and the degree to which reliance on previously observed variation influenced the higher rate of inconclusive results cannot be determined.

Analysis of pathogenic variation in the All of Us Research Program, which integrates data from a diverse cohort to identify genetic differences across ancestries, further highlights the disparities in variant classification across populations. The study examined P/LP variants in a modest number of genes with actionable findings, showing differences as a function of ancestry, with 42% fewer pathogenic variants identified in Latin American versus EUR individuals (1.32% versus 2.26%)[10]. All of Us analyses used Neptune, a system developed for clinical genetic reporting[32]. Neptune relies heavily on variants identified in prior curated data, which will bias the findings in diverse populations. Consistent with this, analyses of the GALA cohort using Neptune show lower rates of findings compared to non-AMR samples. Our results suggest that with a focus on deleterious de novo variation, use of prior results is less necessary, and others have shown that even highly curated variant databases include false-positive findings that can lead to incorrect information to subsequent families[40–43]. Where possible, we recommend minimizing reliance on previously reported pathogenic variants. In addition, to further improve genetic testing results across diverse populations, our results show that it is of key importance to use allele frequency from all relevant populations, as we have done here.

We should, however, recognize the limitations inherent in our study and in any study that focuses on ancestries beyond EUR and a few other commonly characterized populations. For instance, we focused on de novo variants and their interpretation in AMR populations. Variants called de novo in our sample, and within subjects, are likely a mixture of true and false positives. For populations not deeply characterized for genetic variation, it is reasonable to expect elevation in the false-positive rate, simply because we do not know the frequencies of variants therein and which variants are relatively more common. For this reason, more of the variation called de novo is likely to be inherited variation.

At the same time, it is possible that unknown genomic complexity, such as common structural variants[44–46], elevate false negatives within these populations, including genomic variation important for phenotypes like autism, which is another limitation of our study. The combination of these three quantities—true positives, false positives and false negatives—determines the total variation that we observe. Based on our results, which show similar patterns to those observed in EUR studies, we can conclude that the vast majority of our results arise from true positives. Nonetheless, we should not conclude that populations are all the same when it comes to calling de novo variation. Indeed, we can be confident that they are not, given what we know about increased genetic diversity in African populations[47–50] and the impact that cryptic structural variation and singleton events have on the reliability of calling ultra-rare variation. Only through deeper genetic studies can we expect completely comparable results to those of EUR population samples, ameliorating the above issues.

In conclusion, our observations are consistent with the neurobiology of autism being shared across ancestries and provide support for the translatability of autism clinical genetic approaches across ancestries.

## Online content

Any methods, additional references, Nature Portfolio 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/s41591-026-04228-6.

## References


1. Lord, C. et al. Autism spectrum disorder. *Nat. Rev. Dis. Primers* **6**, 5 (2020).
2. Klei, L. et al. Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism* **3**, 9 (2012).
3. Gaugler, T. et al. Most genetic risk for autism resides with common variation. *Nat. Genet.* **46**, 881–885 (2014).
4. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
5. Zhou, X. et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. *Nat. Genet.* **54**, 1305–1319 (2022).
6. 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 (2020).
7. Davidson, B. L. et al. Gene-based therapeutics for rare genetic neurodevelopmental psychiatric disorders. *Mol. Ther.* **30**, 2416–2428 (2022).
8. Fatumo, S. et al. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
9. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
10. Venner, E. et al. The frequency of pathogenic variation in the All of Us cohort reveals ancestry-driven disparities. *Commun. Biol.* **7**, 174 (2024).
11. Abul-Husn, N. S. et al. Molecular diagnostic yield of genome sequencing versus targeted gene panel testing in racially and ethnically diverse pediatric patients. *Genet. Med.* **25**, 100880 (2023).
12. Wright, C. F. et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
13. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
14. Moreno-Estrada, A. et al. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. *Science* **344**, 1280–1285 (2014).
15. Ongaro, L. et al. The genomic impact of European colonization of the Americas. *Curr. Biol.* **29**, 3974–3986 (2019).
16. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
17. DeFelice, M. et al. Blended genome exome (BGE) as a cost efficient alternative to deep whole genomes or arrays. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.03.587209 (2024).
18. Buxbaum, J. D. et al. The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2012).
19. SPARK Consortium. SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).
20. Abul-Husn, N. S. et al. Implementing genomic screening in diverse populations. *Genome Med.* **13**, 17 (2021).
21. Belbin, G. M. et al. Toward a fine-scale population health monitoring system. *Cell* **184**, 2068–2083 (2021).
22. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
23. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2023).

24. Babadi, M. et al. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat. Genet.* **55**, 1589–1597 (2023).
25. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. Preprint at *bioRxiv* https://doi.org/10.1101/148353 (2017).
26. Browning, S. R. et al. Ancestry-specific recent effective population size in the Americas. *PLoS Genet.* **14**, e1007385 (2018).
27. 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).
28. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
29. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
30. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
31. Kopanos, C. et al. VarSome: the human genomic variant search engine. *Bioinformatics* **35**, 1978–1980 (2018).
32. Eric, V. et al. Neptune: an environment for the delivery of genomic medicine. *Genet. Med.* **23**, 1838–1846 (2021).
33. Miller, D. T. et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1381–1390 (2021).
34. Arriaga-MacKenzie, I. S. et al. Summix: a method for detecting and adjusting for population structure in genetic summary data. *Am. J. Hum. Genet.* **108**, 1270–1282 (2021).
35. Gudmundsson, S. et al. Variant interpretation using population databases: lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
36. Hou, K. et al. Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).
37. Saitou, M., Dahl, A., Wang, Q. & Liu, X. Allele frequency impacts the cross-ancestry portability of gene expression prediction in lymphoblastoid cell lines. *Am. J. Hum. Genet.* **111**, 2814–2825 (2024).
38. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).
39. Honorato-Mauer, J. et al. Characterizing features affecting local ancestry inference performance in admixed populations. *Am. J. Hum. Genet.* **112**, 224–234 (2025).
40. Manrai, A. K. et al. Genetic misdiagnoses and the potential for health disparities. *N. Engl. J. Med.* **375**, 655–665 (2016).
41. Ciesielski, T. H., Sirugo, G., Iyengar, S. K. & Williams, S. M. Characterizing the pathogenicity of genetic variants: the consequences of context. *npj Genom. Med.* **9**, 3 (2024).
42. Sharo, A. G., Zou, Y., Adhikari, A. N. & Brenner, S. E. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med.* **15**, 51 (2023).
43. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
44. Jun, G. et al. Structural variation across 138,134 samples in the TOPMed consortium. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-2515453/v1 (2023).
45. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
46. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
47. Yilmaz, F. et al. Genome-wide copy number variations in a large cohort of bantu African children. *BMC Med. Genom.* **14**, 129 (2021).
48. Pereira, L., Mutesa, L., Tindana, P. & Ramsay, M. African genetic diversity and adaptation inform a precision medicine agenda. *Nat. Rev. Genet.* **22**, 284–306 (2021).
49. Gomez, F., Hirbo, J. & Tishkoff, S. A. Genetic variation and adaptation in Africa: implications for human evolution and disease. *Cold Spring Harb. Perspect. Biol.* **6**, a008524 (2014).
50. Yu, N. et al. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161**, 269–274 (2002).
51. Schaaf, C. P. et al. A framework for an evidence-based gene list relevant to autism spectrum disorder. *Nat. Rev. Genet.* **21**, 367–376 (2020).
52. Jay, J. J. & Brouwer, C. Lollipops in the clinic: information dense mutation plots for precision medicine. *PLoS ONE* **11**, e0160519 (2016).




**Marina Natividad Avila[1,2,3,4,5,6], Seulgi Jung[1,2,3,4,5,6], F. Kyle Satterstrom[7,8,9], Jack M. Fu[7,10,11], Tess Levy[1,2,4,6], Laura G. Sloofman[1,2,3,4,5,6], Lambertus Klei[12], Thariana Pichardo[1,2,4,6], Dalia Marquez[1,2], Christine R. Stevens[7,8,9], Caroline M. Cusick[8], Jennifer L. Ames[13], Gabriele S. Campos[14], Hilda Cerros[13], Roberto Chaskel[15,16], Claudia I. S. Costa[14], Michael L. Cuccaro[17,18], Andrea del Pilar Lopez[15], Magdalena Fernandez[16], Eugenio Ferro[16], Liliana Galeano[19], Ana Cristina D. E. S. Girardi[14], Anthony J. Griswold[17,18], Luis C. Hernandez[19], Naila Lourenço[14], Yunin Ludena[20], Diana Núñez-Ríos[21,22], Rosa Oyama[23], Katherine P. Peña[19], Isaac Pessah[20], Rebecca Schmidt[20], Holly M. Sweeney[24], Lizbeth Tolentino[23], Jaqueline Y. T. Wang[14], Lilia Albores-Gallo[25,26], Lisa A. Croen[13,27], Carlos S. Cruz-Fuentes[28],**

**Irva Hertz-Picciotto[20], Alexander Kolevzon[1,2,29], Maria Claudia Lattig[19], Liliana Mayo[23], Maria Rita Passos-Bueno[14], Margaret A. Pericak-Vance[17,18], Paige M. Siper[1,2,6], Flora Tassone[20,30], M. Pilar Trelles[31], GALA Consortium*, The Autism Sequencing Consortium (ASC)*, Michael E. Talkowski[7,8,10,11,32], Mark J. Daly[7,8,9,10,33,34], Behrang Mahjani[1,2,3,6,35,36], Silvia De Rubeis[1,2,4,6,37], Edwin H. Cook[38], Kathryn Roeder[39,40], Catalina Betancur[41], Bernie Devlin[12] & Joseph D. Buxbaum[1,2,3,4,5,6]**

[1]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [9]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [10]Center for Genomic Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [12]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [13]Division of Research, Kaiser Permanente Northern, Pleasanton, CA, USA. [14]Centro de Estudos do Genoma Humano e Células-Tronco, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brasil. [15]Facultad de Medicina, Universidad de los Andes, Bogotá, Colombia. [16]Instituto Colombiano del Sistema Nervioso, Clínica Montserrat, Bogotá, Colombia. [17]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, USA. [18]The Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, USA. [19]Facultad de Ciencias, Universidad de los Andes, Bogotá, Colombia. [20]MIND (Medical Investigation of Neurodevelopmental Disorders) Institute, University of California, Davis, Davis, CA, USA. [21]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [22]National Center of Posttraumatic Stress Disorders, VA CT Healthcare Center, West Haven, CT, USA. [23]Centro Ann Sullivan del Peru, Lima, Peru. [24]Center Ann Sullivan International, Lawrence, KS, USA. [25]Hospital Psiquiátrico Infantil Dr. Juan N. Navarro, Mexico City, Mexico. [26]Universidad Nacional Autónoma de México, Mexico City, Mexico. [27]Kaiser Permanente School of Medicine, Pasadena, CA, USA. [28]Departamento de Genética, Subdirección de Investigaciones Clínicas, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz México, Ciudad de México, Mexico. [29]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [30]Department of Biochemistry and Molecular Medicine, University of California, Davis, School of Medicine, Davis, CA, USA. [31]Psychiatry and Behavioral Sciences, Boston Children's Hospital, Boston, MA, USA. [32]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. [33]Department of Medicine, Harvard Medical School, Boston, MA, USA. [34]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [35]Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [36]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. [37]The Alper Center for Neural Development and Regeneration, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [38]Department of Psychiatry, University of Illinois Chicago, Chicago, IL, USA. [39]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA. [40]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA. [41]Sorbonne Université, INSERM, CNRS, Institut de Biologie Paris Seine, Center for Neuroscience at Sorbonne Université, Paris, France. *Lists of authors and their affiliations appear at the end of the paper. e-mail: joseph.buxbaum@mssm.edu

## GALA Consortium

**Lilia Albores-Gallo[25,26], Jennifer L. Ames[13], Catalina Betancur[41], Joseph D. Buxbaum[1,2,3,4,5,6], Gabriele S. Campos[14], Hilda Cerros[13], Roberto Chaskel[15,16], Edwin H. Cook[38], Claudia I. S. Costa[14], Lisa A. Croen[13,27], Carlos S. Cruz-Fuentes[28], Michael L. Cuccaro[17,18], Silvia De Rubeis[1,2,4,6,37], Bernie Devlin[12], Magdalena Fernandez[16], Eugenio Ferro[16], Jennifer Foss-Feig[1,2,6], Liliana Galeano[19], Ana Cristina D. E. S. Girardi[14], Anthony J. Griswold[17,18], Erina Hara[1,2], Danielle Halpern[1,2], Luis C. Hernandez[19], Irva Hertz-Picciotto[20], Seulgi Jung[1,2,3,4,5,6], Lambertus Klei[12], Alexander Kolevzon[1,2,29], Maria Claudia Lattig[19], Tess Levy[1,2,4,6], Yi Li[1,2], Andrea del Pilar Lopez[15], Naila Lourenço[14], Yunin Ludena[20], Behrang Mahjani[1,2,3,6,35,36], Dalia Marquez[1,2], Liliana Mayo[23], Marina Natividad Avila[1,2,3,4,5,6], Diana Núñez-Ríos[21,22], Rosa Oyama[23], Maria Rita Passos-Bueno[14], Katherine P. Peña[19], Margaret A. Pericak-Vance[17,18], Isaac Pessah[20], Thariana Pichardo[1,2,4,6], Kathryn Roeder[39,40], Catherine Sancimino[1,2], Rebecca Schmidt[20], Paige M. Siper[1,2,6], Laura G. Sloofman[1,2,3,4,5,6], Renee Soufer[1,2], Holly M. Sweeney[24], Flora Tassone[20,30], Lizbeth Tolentino[23], M. Pilar Trelles[31], Jaqueline Y. T. Wang[14] & Jessica Zweifach[1,2]**

## The Autism Sequencing Consortium (ASC)

**Branko Aleksic[42], Mykyta Artomov[7,8,11,43], Mafalda Barbosa[3,6], Elisa Benetti[44,45], Catalina Betancur[41], Monica Biscaldi-Schafer[46], Anders D. Børglum[47,48,49,50], Harrison Brand[7,11,43,51], Alfredo Brusco[52,53], Joseph D. Buxbaum[1,2,3,4,5,6], Gabriele S. Campos[14], Simona Cardaropoli[54], Diana Carli[54], Angel Carracedo[55,56], Marcus C. Y. Chan[57], Andreas G. Chiocchetti[42], Brian H. Y. Chung[57], Brett Collins[1,2,6], Ryan L. Collins[7,11,32,43], Edwin H. Cook[38], Hilary Coon[58,59], Claudia I. S. Costa[14], Michael L. Cuccaro[17,18], David J. Cutler[60], Mark J. Daly[7,8,9,10,33,34], Silvia De Rubeis[1,2,4,6,37], Bernie Devlin[12], Ryan N. Doan[61], Enrico Domenici[62], Shan Dong[63], Chiara Fallerini[44,45], Magdalena Fernandez[16], Montserrat Fernández-Prieto[55,64], Giovanni Battista Ferrero[54], Eugenio Ferro[16], Jennifer Foss-Feig[1,2,6], Christine M. Freitag[42], Jack M. Fu[7,10,11], Liliana Galeano[19], J. Jay Gargus[65], Sherif Gerges[7,8,11,43], Elisa Giorgio[52], Ana Cristina D. E. S. Girardi[14], Stephen Guter[66], Emily Hansen-Kiss[67], Erina Hara[1,2], Gail E. Herman[68], Luis C. Hernandez[19], Irva Hertz-Picciotto[20], David M. Hougaard[46,69], Christina M. Hultman[70], Suma Jacob[66],**

**Miia Kaartinen[71], Lambertus Klei[12], Alexander Kolevzon[1,2,29], Itaru Kushima[47,72], Maria Claudia Lattig[19], So Lun Lee[57], Terho Lehtimäki[73], Lindsay Liang[63], Carla Lintas[74], Alicia Ljungdahl[63], Andrea del Pilar Lopez[15], Caterina Lo Rizzo[44,45], Yunin Ludena[20], Patricia Maciel[75], Behrang Mahjani[1,2,3,6,35,36], Nell Maltman[66], Marianna Manara[45,76], Dara S. Manoach[77], Gal Meiri[78,79], Idan Menashe[80,81], Judith Miller[82,83], Nancy Minshew[12], Matthew Mosconi[84], Marina Natividad Avila[1,2,3,4,5,6], Rachel Nguyen[65], Norio Ozaki[47,85], Aarno Palotie[7,9,34,86], Mara Parellada[87], Maria Rita Passos-Bueno[14], Lisa Pavinato[52], Katherine P. Peña[19], Minshi Peng[88], Margaret Pericak-Vance[17,18], Antonio M. Persico[89], Isaac N. Pessah[20], Thariana Pichardo[1,2,4,6], Kaija Puura[71], Abraham Reichenberg[1,2,6,90], Alessandra Renieri[44,45,76], Kathryn Roeder[39,40], Catherine Sancimino[1,2], Stephan J. Sanders[91,92,93], Sven Sandin[1,2,70], F. Kyle Satterstrom[7,8,9], Stephen W. Scherer[94,95], Sabine Schlitt[42], Rebecca J. Schmidt[20], Lauren Schmitt[66], Katja Schneider-Momm[42], Paige M. Siper[1,2,6], Laura Sloofman[1,2,3,4,5,6], Moyra Smith[65], Renee Soufer[1,2], Christine R. Stevens[7,8,9], Pål Suren[96], James S. Sutcliffe[97,98], John A. Sweeney[99], Michael E. Talkowski[7,8,10,11,32], Flora Tassone[20,30], Karoline Teufel[42], Elisabetta Trabetti[100], Slavica Trajkova[52], M. Pilar Trelles[31], Brie Wamsley[101], Jaqueline Y. T. Wang[14], Lauren A. Weiss[63], Mullin H. C. Yu[57] & Ryan Yuen[94]**

[42]Department of Psychiatry, Graduate School of Medicine, Nagoya University, Nagoya, Japan. [43]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [44]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Siena, Italy. [45]Medical Genetics, University of Siena, Siena, Italy. [46]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe University Frankfurt, Frankfurt, Germany. [47]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Aarhus, Denmark. [48]Department of Biomedicine—Human Genetics, Aarhus University, Aarhus, Denmark. [49]Center for Genomics and Personalized Medicine, Aarhus, Denmark. [50]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [51]Pediatric Surgical Research Laboratories, Department of Surgery, Massachusetts General Hospital, Boston, MA, USA. [52]Department of Medical Sciences, University of Torino, Turin, Italy. [53]Medical Genetics Unit, 'Città della Salute e della Scienza' University Hospital, Turin, Italy. [54]Department of Public Health and Pediatrics, University of Torino, Turin, Italy. [55]Grupo de Medicina Xenómica, Centro de Investigación en Red de Enfermedades Raras (CIBERER), CIMUS, Universidade de Santiago de Compostela, Santiago de Compostela, Spain. [56]Fundación Pública Galega de Medicina Xenómica, Servicio Galego de Saúde (SERGAS), Santiago de Compostela, Spain. [57]Department of Pediatrics and Adolescent Medicine, Duchess of Kent Children's Hospital, The University of Hong Kong, Hong Kong Special Administrative Region, Hong Kong, China. [58]Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA. [59]Department of Psychiatry, Huntsman Mental Health Institute, University of Utah, Salt Lake City, UT, USA. [60]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA. [61]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [62]Department of Cellular, Computational and Integrative Biology, University of Trento, Trento, Italy. [63]Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [64]Neurogenetics group, Instituto de Investigación Sanitaria de Santiago (IDIS-SERGAS), Santiago de Compostela, Spain. [65]Center for Autism Research and Translation, University of California, Irvine, Irvine, CA, USA. [66]Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, IL, USA. [67]Department of Diagnostic and Biomedical Sciences, The University of Texas Health Science Center at Houston, School of Dentistry, Houston, TX, USA. [68]The Research Institute at Nationwide Children's Hospital, Columbus, OH, USA. [69]Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark. [70]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [71]Department of Child Psychiatry, Tampere University and Tampere University Hospital, Tampere, Finland. [72]Medical Genomics Center, Nagoya University Hospital, Nagoya, Japan. [73]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [74]Service for Neurodevelopmental Disorders, University Campus Bio-medico of Rome, Rome, Italy. [75]Life and Health Sciences Research Institute, School of Medicine, University of Minho, Braga, Portugal. [76]Genetica Medica, Azienda Ospedaliera Universitaria Senese, Siena, Italy. [77]Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [78]The Azrieli National Center for Autism and Neurodevelopment Research, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [79]Pre-School Psychiatry Unit, Soroka University Medical Center, Be'er-Sheva, Israel. [80]Department of Public Health, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [81]National Autism Research Center of Israel, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. [82]Children's Center for Autism Research and Training, University of Kansas, Lawrence, KS, USA. [83]Department of Psychiatry, University of Utah, Salt Lake City, UT, USA. [84]Life Span Institute and Kansas Center for Autism Research and Training, University of Kansas, Lawrence, KS, USA. [85]Institute for Glyco-core Research (iGCORE), Nagoya University, Nagoya, Japan. [86]Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [87]Department of Child and Adolescent Psychiatry, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, School of Medicine Complutense University, Madrid, Spain. [88]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, USA. [89]Interdepartmental Program 'Autism 0-90', 'Gaetano Martino' University Hospital, University of Messina, Messina, Italy. [90]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [91]Institute of Developmental and Regenerative Medicine, Department of Paediatrics, University of Oxford, Oxford, UK. [92]Department of Psychiatry and Behavioral Sciences, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [93]New York Genome Center, New York, NY, USA. [94]Program in Genetics and Genome Biology, The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada. [95]Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada. [96]Norwegian Institute of Public Health, Oslo, Norway. [97]Department of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt University School of Medicine, Nashville, TN, USA. [98]Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN, USA. [99]Department of Psychiatry, University of Cincinnati, Cincinnati, OH, USA. [100]Department of Neurosciences, Biomedicine and Movement Sciences, Section of Biology and Genetics, University of Verona, Verona, Italy. [101]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.

## Methods

### Cohort description

GALA comprises multiple sites from North, Central and South America recruiting AMR participants for studies on the genetic architecture of autism. Study procedures were approved by the institutional review board (IRB) of the Program for the Protection of Human Subjects at Mount Sinai (no. 16-01262). Informed consent was obtained from the parents or legal guardians of all study participants.

Study procedures for participant enrollment were approved by the Program for the Protection of Human Subjects at Mount Sinai (no. 16-01262 for the Seaver Center at Mount Sinai, São Paulo, Brazil, and Bogotá, Colombia; and no. 21-00039 for Peru), the University of California, Davis IRB (no. 226028-22) and the University of Miami IRB (no. 20070193). Two cohorts were collected previously: study procedures for participant enrollment in Costa Rica were approved under the guidelines of the Ministry of Health of Costa Rica, the Ethical Committee of the National Children's Hospital in San Jose and the IRB at Mount Sinai, as described previously[53,54]; and The Autism Simplex Collection (TASC), which included an estimated 12% of individuals of Latin American ancestry, was recruited across 13 sites in North America and Europe, as described previously[55], with local IRB oversight and all consents reviewed before depositing biospecimens and data to the National Institutes of Health repository.

For clarity, we use 'ASD' to refer to individuals who received a clinical diagnosis according to the procedure outlined below and 'autism' elsewhere. ASD diagnoses are based on expert clinical evaluations using *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edition (DSM-5) criteria, incorporating all available data, including standardized assessments. Participants can be any age. Individuals with a known genetic condition (for example, fragile X syndrome) are excluded from analyses. Once a diagnosis of ASD is confirmed, the individual and their parents contribute a sample (blood or saliva) for genetic analyses. If both parents are not available, collection of other biological family members is encouraged (siblings, grandparents, etc.). Participating sites generally also collect additional clinical and family history information.

### Description of GALA sites

**New York, USA.** The Seaver Autism Center for Research and Treatment at the Icahn School of Medicine at Mount Sinai, located in New York City, is the main coordinating site within the GALA Consortium. AMR individuals make up almost 30% of the population of New York City. Affected individuals undergo a full diagnostic ASD workup and receive additional assessments, including a cognitive test, adaptive behavior measure, medical checklist and behavioral checklists. Participating families receive $100 USD in compensation.

**São Paulo, Brazil.** The Human Genome and Stem Cell Research Center (HUG-CELL) at the Universidade de São Paulo in Brazil has over 20 years of experience in clinical and molecular research in autism, with more than 2,000 families seen. Brazil has a multiethnic admixed population, including African and Amerindian ancestry[56]. The HUG-CELL conducts research in human and medical genetics of rare diseases, providing genetic counseling services and genetic tests for the population. A team of psychiatrists, psychologists and neurologists completes a formal ASD diagnostic workup prior to obtaining samples for genetic testing from the individual and their family members. Financial compensation for participation is not permitted at this site; however, individuals who meet clinical criteria are offered free fragile X testing.

**Bogotá, Colombia.** The Centro de Investigaciones Genéticas en Enfermedades Humanas (CIGEn) at the Universidad de los Andes in Bogotá, Colombia, in close collaboration with the Instituto Colombiano del Sistema Nervioso, Clínica Montserrat, focuses on unraveling the prevalence and characteristics of autism within the Colombian population. Through ASD referrals, the impact of CIGEn extends beyond Bogotá, reaching out to other cities throughout Colombia (Medellín, Cali, Armenia, Pereira, Bucaramanga, Cartagena, Barranquilla and Santa Marta), with the aim of including families from diverse backgrounds. Financial compensation is not offered for participation.

**Mexico City, Mexico.** The Children's Psychiatric Hospital 'Juan N. Navarro' (HPIJNN), which is part of the Psychiatric Care Services of the Mexican Government's Ministry of Health, provides professional care for minors with mental health, psychiatric and behavioral problems. As the largest teaching center in child and adolescent psychiatry in Mexico, it performs diverse biomedical and clinical research activities. One of the main lines of research focuses on autism, in collaboration with the Genetics Department at the National Institute of Psychiatry Ramón de la Fuente Muñíz (INPRFM). The samples from Mexico are being sequenced and were not included in the current analyses. Financial compensation is not provided at this site, in accordance with ethics committee requirements.

**Lima, Peru.** The Centro Ann Sullivan del Perú is a non-profit center in Lima, Peru, that serves individuals with varying abilities and their families. The center specializes in helping individuals with ASD. GALA investigators from the Seaver Autism Center (M.P.T. and A.K.) traveled to Lima to perform 40 psychiatric evaluations, aid in ASD diagnostics and collect blood samples from individuals with ASD and their families. Behavioral surveys were carried out for all participants, and ASD and attention-deficit/hyperactivity disorder diagnoses were made using DSM-5 criteria. Financial compensation was not offered; instead, participating individuals received their clinical evaluation results.

**California, USA (CHARGE).** The Childhood Autism Risks from Genetics and the Environment (CHARGE) cohort is a population-based case–control study collected in California at the University of California, Davis, Center for Children's Environmental Health laboratories with the intent of addressing the impact of environmental exposures on risk[57].

**Florida, USA.** The John P. Hussman Institute for Human Genomics at the University of Miami, located in Miami, Florida, recruits families through clinical referrals and lay organizations, providing services to families with ASD. Upwards of 70% of the Miami population identifies as AMR. The diagnostic workup included the Autism Diagnostic Interview-Revised (ADI-R) and assessment of adaptive behavior. Discrepancies between ADI-R and clinical findings were resolved using additional clinical measures, including the Autism Diagnostic Observation Schedule (ADOS).

**Central Valley, Costa Rica.** The founder population of the Central Valley of Costa Rica (CVCR) originated at the end of the 16th century from the intermarriage of 86 Spanish families and Indigenous Americans. The population was geographically isolated until the late 19th century; therefore, the current inhabitants are estimated to descend from fewer than 1,000 founders[58]. A genetic study on autism in the CVCR was initiated in 2003, and affected individuals were ascertained using the translated Spanish versions of the ADI-R and the ADOS as well as assessment of intellectual abilities and adaptive behavior[53].

**USA and Europe (TASC).** TASC was a collaboration among 13 sites in North America and Western Europe funded by the National Alliance for Autism Research, now Autism Speaks, and the National Institute of Mental Health. As detailed previously[55], more than 1,700 individuals with ASD confirmed with extensive prospective assessment, as well as additional family members including parents, completed this study. Individuals within this study were sequenced, and those who were of AMR ancestry were included in these analyses.

**California, USA (Kaiser Permanente).** The Autism Research Program (ARP) at the Kaiser Permanente Northern California (KPNC) Division of Research was established in 2002 by Senior Research Scientist Lisa Croen. The program focuses on research identifying genetic and environmental factors associated with autism and understanding patterns of detection, diagnosis and utilization of health services for individuals with ASD across the lifespan. The ARP created the Autism Family Biobank, a repository including genetic, medical and environmental information from more than 1,000 individuals with ASD and their two biological parents, who donated blood or saliva between 2015 and 2017. This collection is representative of the diverse population served by KPNC, an integrated healthcare system. The samples from Kaiser Permanente are being sequenced and were not included in the current analyses. Participants receive $15 USD per biospecimen, and families receive an additional $15 USD upon completion of the parent surveys.

### Ancestry determination and sample-level quality control

Latin American samples analyzed in the current freeze include (1) GALA participants (some published in Fu et al.[4]); (2) non-overlapping AMR samples in the ASC and SPARK[19] reported in Fu et al.[4]; and (3) additional AMR samples from the new release of SPARK (iWESv2). The current freeze includes trio data from 14,359 AMR samples, including 4,450 affected individuals (609 from GALA and 3,841 from ASC and the SPARK releases) and 1,459 typically developing siblings and case–control data from 267 cases and 801 controls.

To assign ancestry to each case, we followed an approach modeled after the pipeline used by gnomAD (https://gnomad.broadinstitute.org/news/2021-09-using-the-gnomad-ancestry-principal-components-analysis-loadings-and-random-forest-classifier-on-your-dataset/). Specifically, each of three jointly called datasets, derived from unpublished GALA sequencing, Fu et al. and SPARK (iWESv2), was merged with the Human Genome Diversity Project (HGDP) + 1000 Genomes Project (1KG) subset of gnomAD[22], and principal component analysis (PCA) was performed in the joint dataset after they had been restricted to 5,000 ancestry-informative single-nucleotide polymorphisms[59]. A random forest classifier was trained on the HGDP + 1KG reference samples using the first 10 principal components and used to assign superpopulation/continental ancestry to individuals in our dataset. AMR ancestry classification was based on the predicted ancestry label assigned by the random forest model. Non-AMR cases included any individuals with ASD in ASC or SPARK releases who did not meet our criteria for genetically inferred AMR ancestry (28,818 parents, 13,030 probands and 4,749 typically developing siblings).

Hail 0.2 was used to process the SPARK (iWESv2) and unpublished GALA joint-genotyped variant call files (VCFs). Multiallelic sites were split; variants were annotated using the Variant Effect Predictor (VEP)[60]; and low-complexity regions (https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs38.bed.gz) were removed. Hail's pc_relate() function was used to confirm reported pedigrees and identify duplicate samples within and between datasets, which were removed. Sex was imputed using the impute_sex() function, and genotype filters were applied as described in previous methodology[6] to generate working datasets (Extended Data Fig. 1).

### De novo variants

Previously published de novo calls were extracted from Supplementary Table 20 from Fu et al.[4]. For the unpublished GALA and SPARK (iWESv2) datasets, de novo variants were called using the my_de_novo_v16() function (https://discuss.hail.is/t/de-novo-calls-on-hemizygous-x-variants/2357/19) with variant frequencies from the non-neuro subset of gnomAD exomes version 2.1.1 as priors. Potential de novo variants were dropped if they were present at a frequency greater than 0.1% within the non-neuro subset of gnomAD version 2.1.1, gnomAD version 3.1.2, in any subpopulation of these gnomAD datasets or the dataset in which they were called. Variants were further excluded if they had 'ExcessHet' in the Filters field, exhibited a proband allele balance < 0.3 or demonstrated a depth ratio < 0.3. Only 'HIGH'-confidence or 'MEDIUM'-confidence variants were kept, with the MEDIUM-confidence calls limited to a maximum allele count in the dataset of 1. A single variant per person per gene was chosen, giving preference to variants with more damaging consequences. Samples were finally excluded if the count of coding de novo variants was significantly greater than expected.

### Inherited variants

Starting with the same working datasets as for de novo calling, counts of transmitted and non-transmitted alleles were generated using Hail's transmission_disequilibrium_test() function. Variants were filtered out if they were marked 'ExcessHet' by GATK4 or had allele frequencies greater than 0.01% within their own dataset, within the non-neuro subset of gnomAD version 2.1.1, gnomAD version 3.1.2 or within any subpopulation of these gnomAD datasets. Variants with an allele count > 6 in the total parents of the dataset were excluded as well. Hard filtering was applied according to GATK recommendations (https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants). Final counts of transmitted and non-transmitted alleles were produced for PTV, MisB, MisA ($1 \leq MPC < 2$) and synonymous variants.

### Case–control variants

Probands within incomplete trios were identified from the ASC and GALA cohorts and matched using the top 10 principal components ('Ancestry determination and sample-level quality control') with non-psychiatric, unrelated controls from BioMe at a ratio of three controls to one case (3:1). Incomplete trios from SPARK (iWESv2) were removed. To ensure genome build standardization between these two cohorts, CRAM files from ASD cases were unmapped using GATK4 (ref. 61) and then remapped to a different version of the hg38 reference genome (https://biobank.ndph.ox.ac.uk/ukb/refer.cgi?id=838) using GATK3.5. Single-nucleotide variants (SNVs) and insertions/deletions (indels) were joint-genotyped across cases using the Haplotypecaller of GATK4. Like for the trio dataset processing, Hail 0.2 (https://hail.is) was used to process the joint-genotyped VCF file. The identity_by_descent() function of Hail was used to test for relatedness, which resulted in the removal of 13 cases. Sex was imputed for every sample using the impute_sex() function of Hail and cross-checked with metadata provided by all sites to ensure sample concordance.

As was done for the previous datasets[4], multiallelic sites were split, variants were annotated using the VEP and low-complexity regions were removed. Variants were removed if they had an allele count ≥ 2 in the entire case–control dataset as well as an allele count ≥ 5 in the non-psychiatric subset of gnomAD version 2.1.1. Genotype calls were filtered to genotype quality > 25 and allele balance > 0.3. For case–control coverage harmonization, variants in high coverage, defined as a call rate ≥ 90%, were kept. To perform case–control matching, we excluded one case that was an outlier in the distribution of the number of synonymous variants. Finally, 267 cases were matched to 801 controls by sex and the first 10 principal components using the match_on function of the R package optmatch[62].

### CNV analysis

De novo CNVs called in Fu et al.[4] coming from AMR samples were extracted (1,861 probands and 680 unaffected siblings). Trio and case–control datasets were analyzed separately, and GATK-gCNV[24] was used to detect CNVs. First, raw CRAM files were compressed into read counts that covered the annotated exons to serve as input data. Then, a PCA-based approach that combines density and distance-based clustering was employed on the observed read counts to organize batches of samples for parallel processing. GATK-gCNV was run on cohort mode analysis for 200 samples within the cluster identified through PCA, and the remaining samples were subjected to GATK-gCNV analysis using

the case mode, with models specific to the cohort (368 probands and 29 typically developing siblings). For quality control, CNV calls were processed according to Fu et al.[4] methodology; CNVs were retained if they had an allele frequency < 1% that spanned more than two captured exons. For homozygous deletions, the quality score threshold was set to the lesser of 400 or 10 times the number of intervals. For heterozygous deletions, the quality score threshold was set to the lesser of 100 or 10 times the number of intervals. For duplications, the quality score threshold was set to the lesser of 50 or four times the number of intervals. For sample-level quality control, samples were retained if the number of raw, autosomal CNV calls detected by GATK-gCNV did not exceed 200 and if the number of calls with quality score ≥ 20 did not exceed 35. After quality control, 291 probands, 25 typically developing siblings, 209 cases and 735 controls remained.

A gene was considered impacted by a deletion if at least 10% of its non-redundant exons were overlapped by the deletion. For a duplication, a gene was considered impacted if at least 75% of its non-redundant exons were overlapped. Additionally, CNVs were annotated against a list of 79 curated genomic disorder loci (see Supplementary Table 10 in Fu et al.[4]), and a CNV call was classified as a genomic disorder CNV if it shared at least 50% reciprocal overlap with an annotated genomic disorder.

### Genetic association analyses

TADA[4,27] was performed for three types of inheritance classes: de novo (PTV, MisB, MisA, deletion (DEL) and duplication (DUP)), inherited (PTV, MisB and MisA) and case–control (PTV, MisB, MisA, DEL and DUP) variation. CNVs resulting from non-allelic homologous recombination (NAHR) were excluded, and only CNVs impacting fewer than nine constrained genes were retained (LOEUF < 0.6) (Supplementary Tables 13–19).

Bayes factors were constructed separately for each variant class (PTV, MisA, MisB, DEL and DUP) as described, accounting for sample size and directly using relative risk priors from Fu et al. directly (see Supplementary Table 8 in Fu et al.[4]). Previously published mutation rates were adjusted to align with the observed variant counts in unaffected siblings for each variant type in the dataset[4].

### Expected versus observed mutations in GALA

As noted in the main text, for top genes in GALA that were also significant in $Fu_{COMP}$, we compared the observed numbers of variants in the GALA cohort with the expected number of variants derived from TADA analysis in $Fu_{COMP}$. Although observed and expected counts may vary at the individual gene level, as expected for ultra-rare events, the overall observed and expected totals across all genes are well matched, supporting the consistency of signal with expectation (Extended Data Table 1).

### Clinical genetics analyses

In addition to VarSome described in the main text, we also ran Neptune[32], which uses databases of previously identified variants to call P/LP variants in a set of target genes; we took a similar approach to a recent study[10] carried out in the All Of Us Research Program by focusing on 73 actionable ACMG genes[63]. Of the 12,162 variants in these genes among the 4,450 family-based AMR cases, Neptune provided a classification for 8,501 (69.9%); this compares to 28,262 variants among the 13,030 non-AMR family-based cases, of which 20,750 (73.4%) were classified by Neptune. In AMR participants, 136 variants were classified as P/LP, representing 1.12% (136/12,162) of all variants in these genes and 1.60% (136/8,501) of all classified variants. In non-AMR participants, 344 variants were classified as P/LP, representing 1.22% (344/28,262) of all variants and 1.66% (344/20,750) of all classified variants. Examining the results from the perspective of the participants, in AMR we observed 2.73 variants in these genes per individual, of which 1.91 per individual could be classified by Neptune, and 0.031 per individual were classified as P/LP. The corresponding numbers were 2.17 variants, 1.59 Neptune classified variants and 0.026 P/LP variants per non-AMR individual. The results show that, on the variant level, the differences in AMR versus non-AMR participants trace in part to a reduced ability of Neptune to classify non-AMR variants (Extended Data Fig. 6 and Supplementary Table 11). However, as also noted above, there are more variants per AMR participant (both total and Neptune classified), leading to an apparent lessening of impact in terms of P/LP variants per individual.

### ACMG interpretation of variants

As noted above, for genetic association analyses, the TADA framework was limited to autosomal genes with available mutation rates and LOEUF scores ($n$ = 18,128 genes) and considered only missense variants with an MPC score ≥ 1. By contrast, the clinical interpretation of variants included all autosomal or X-linked protein-truncating, synonymous and missense variants, regardless of gene annotation. In addition to applying the allele frequency cutoff of 0.1% ('De novo variants'), X-linked variants were subjected to an allele frequency cutoff of 0.1% in the male non-psychiatric subsets of gnomAD versions 2.1.1 and 3.1.2 and their subpopulations. This resulted in 20,571 de novo variants being included for clinical genetics annotation. Inherited variant analysis was restricted to a list of well-established X-linked genes implicated in autism and/or intellectual disability (Supplementary Table 9) and subjected to the same allele frequency cutoff.

The commercially available VarSome package[31] was used to evaluate the clinical impact of both de novo variants and X-linked inherited variation in the selected genes. Given the large number of variants, a batch environment was used, which limited the parameters that could be optimized for each gene. Additionally, as ACMG guidelines[30] consider patient phenotype, the focus was placed on genes for which there was a reported relationship with an autism phenotype (Autism Spectrum Disorder, Autism and Autistic Behavior) and/or with a broader NDD phenotype (including the three autism terms as well as Intellectual Disability, Global Developmental Delay, Seizure, Epileptic Encephalopathy and Complex Neurodevelopmental Disorder), without knowing the full spectrum of non-autism phenotypes in the participants. Hence, the results presented here (Supplementary Tables 10–12), although based on a more transparent algorithm, should not be considered fully compliant with ACMG classification guidelines.

The api.batch_lookup function in VarSome was used to obtain germline variant-level information related to ACMG classification, nucleotide substitution and amino acid substitution, along with pathogenicity predictions. When possible, transcripts with the most severe coding impact were selected. Otherwise, the MANE Select transcript, longest canonical transcript, MANE Plus transcript, longest transcript or RefSeq transcript was chosen in that order by default.

For de novo variation, variant lists containing unique sets of variants found in each sex and zygosity were annotated. Inheritance in VarSome was set to 'Confirmed De Novo'. Output from each list was returned in separate JSON files, which were then read into R for downstream processing into tab-separated tables. Inherited variation was examined in a similar manner; however, inheritance was set to the parent of origin of the variant.

To extend these analyses further, we used Neptune[32], examining 73 ACMG actionable genes analyzed in All Of Us[10]. The VIP database used for annotation in Neptune was downloaded from https://gitlab.com/bcm-hgsc/neptune in VCF format, and all variants were lifted over[63] from GRCh37 to GRCh38. Clinical significance annotations were parsed from the INFO field, and variants classified as Pathogenic/Likely Pathogenic, Uncertain significance and Benign/Likely Benign were noted. All rare variants in probands, regardless of mode of inheritance, were used in these analyses. Of the 73 genes, Venner et al.[10] annotated only biallelic variants as P/LP in three recessive genes (*MUTYH*, *ATP3B* and *KCNQ1*) and only a specific variant as P/LP in *HFE*;

we did not observe P/LP variants in these four genes, so no additional corrections were made.

### Inclusion and ethics statement

This study was conducted in accordance with Nature Portfolio's guidelines on inclusion and ethics in global research. The research was designed to include participants of diverse ancestries, with the goal of improving representation in autism genetics research. Study protocols were approved by the IRBs at all participating sites, including the Program for the Protection of Human Subjects at the Icahn School of Medicine at Mount Sinai (GCO no. 14-1082(0001)) as well as the local IRBs in Brazil, Colombia, Peru, Mexico, Kaiser Permanente and the CHARGE study (see 'Cohort description'). Written informed consent was obtained from all participants or from parents or legal guardians where necessary. Data collection adhered to relevant ethical and cultural standards, and compensation for participation varied by site as described above. Collaborations between institutions in the United States and Latin America were established to ensure equitable contributions across sites. Local investigators in Brazil, Colombia, Mexico and Perú were involved in data collection and authorship.

Sex was recorded based on self-report at enrollment and confirmed with genetic information. Both male and female participants were included; however, sex-stratified analyses were not conducted, as the primary focus of this study was on de novo and rare variant burden across ancestry groups rather than sex differences. Participant ages varied by cohort, with probands typically enrolled during childhood or adolescence and parents as adults.

### Statistics and reproducibility

All statistical analyses were performed using R (version 4.3.3), Hail (version 2.0) and Python (version 3.8). Statistical methods are described in detail in the relevant sections of Methods. Two-sided tests were used throughout unless otherwise specified. Multiple hypothesis testing was corrected using the Benjamini–Hochberg FDR procedure or Bonferroni correction as appropriate. Sample sizes were determined by the number of available participants meeting inclusion criteria in the ASC, GALA and SPARK cohorts; no statistical method was used to predetermine sample size. All available samples passing relatedness and quality control thresholds were included in the analyses. No data were otherwise excluded from the analyses.

Because this study involved secondary analysis of existing human genomic data, randomization and blinding were not applicable. The investigators were not blinded to sample status during analyses. Scripts for computational analyses performed were deposited in a GitHub repository (https://github.com/buxbaum-lab/GALA) to ensure reproducibility. Key results were independently replicated using validation datasets as described.

### Reporting summary

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

## Data availability

Sequencing data for the ASC and GALA samples are available through controlled access via the Database of Genotypes and Phenotypes (accession number phs002502) and on the National Human Genome Research Institute Genomic Data Science Analysis, Visualization and Informatics Lab-space (AnVIL) under accession number phs002502.v1.p1 (https://anvilproject.org/data). SPARK phenotype and sequencing data are available to authorized users through the SFARI Base (https://www.sfari.org/resource/sfari-base/).

Individual-level data from the ASC and GALA cohorts are not publicly available due to participant privacy restrictions. Researchers may request access by contacting J.D.B. (joseph.buxbaum@mssm.edu). All requests will be reviewed by the Mount Sinai Institutional Data Access Committee to ensure compliance with participant consent and IRB protocols. Reasonable requests will receive a response within 2–4 weeks. Summary variant counts, gene-level burden statistics and figure source data are available in the accompanying Supplementary Tables and at https://github.com/buxbaum-lab/GALA.

## Code availability

All software used in this study is publicly available at the cited references. The R code used to generate the TADA analysis and figures is available under the MIT license at https://github.com/buxbaum-lab/GALA.

## References


53. McInnes, L. A. et al. A genetic study of autism in Costa Rica: multiple variables affecting IQ scores observed in a preliminary sample of autistic cases. *BMC Psychiatry* **5**, 15 (2005).
54. McInnes, L. A. et al. The NRG1 exon 11 missense variant is not associated with autism in the Central Valley of Costa Rica. *BMC Psychiatry* **7**, 21 (2007).
55. Buxbaum, J. et al. The Autism Simplex Collection: an international, expertly phenotyped autism sample for genetic and phenotypic analyses. *Mol. Autism* **5**, 34 (2014).
56. Naslavsky, M. S. et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum. Mutat.* **38**, 751–763 (2017).
57. Hertz-Picciotto, I. et al. The CHARGE study: an epidemiologic investigation of genetic and environmental factors contributing to autism. *Environ. Health Perspect.* **114**, 1119–1125 (2006).
58. Mathews, C. A. et al. Genetic studies of neuropsychiatric disorders in Costa Rica: a model for the use of isolated populations. *Psychiatr. Genet.* **14**, 13–23 (2004).
59. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
60. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
61. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
62. Hansen, B. B. & Olsen Klopfer, S. Optimal full matching and related designs via network flows. *J. Comput. Graph. Stat.* **15**, 609–627 (2006).
63. Miller, D. T. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **23**, 1391–1398 (2021).


## Acknowledgements


GALA is currently supported by the National Institutes of Health (grant MH128813, J.D.B.), the Seaver Autism Center for Research and Treatment and the SWT and Seaver Foundations. GALA originated with sites from, and with support of, the ASC (MH129724, J.D.B.; MH129722, M.D.; MH129725, K.R; MH129751, S.S.; and prior ASC funding—for example, MH100233 and MH111661). ASC sites continue to support analyses of GALA studies, with additional analyses supported by MH128813. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by Clinical and Translational Science Awards grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this paper was also supported by the Office of Research Infrastructure of the National Institutes of Health under award numbers S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study makes use

of data generated by the DECIPHER community. A full list of centers that contributed to the generation of the data is available from https://deciphergenomics.org/about/stats and via email from contact@deciphergenomics.org. DECIPHER is hosted by the EMBL-EBI, and funding for the DECIPHER project was provided by the Wellcome Trust (grant no. WT223718/Z/21/Z).

## Author contributions

K.R., B.D., C.B. and J.D.B. conceived and designed the study. T.L., T.P., C.R.S., C.M.C., J.L.A., G.S.C., H.C., R.C., C.I.S.C., M.L.C., A.D.P.L., M.F., E.F., L.G., A.C.D.E.S.G., A.J.G., L.C.H., N.L., Y.L., D.N.-R., R.O., K.P.P., I.P., R.S., H.M.S., L.T., J.Y.T.W., L.A.-G., L.A.C., C.S.C.-F., I.H.-P., A.K., M.C.L., L.M., M.R.P.-B., M.A.P.-V., P.S., F.T., M.P.T., M.E.T., M.J.D. and J.D.B. contributed samples and generated data. J.D.B., C.B., B.D., B.M., S.D.R., L.K., L.S., J.M.F., F.K.S., S.J. and M.N.A. developed methodology and performed data analyses. J.D.B., B.D., C.B., E.H.C., T.L., L.S. and M.N.A. drafted and revised the paper. All authors reviewed and approved the final version of the paper. J.D.B. supervised the study.

## Competing interests

L.A.-G. is the main author of the CRIDI-ASD interview; she teaches the training course for the aforementioned instrument and receives payment for the training. The other authors declare no conflicts of interest.

## Additional information

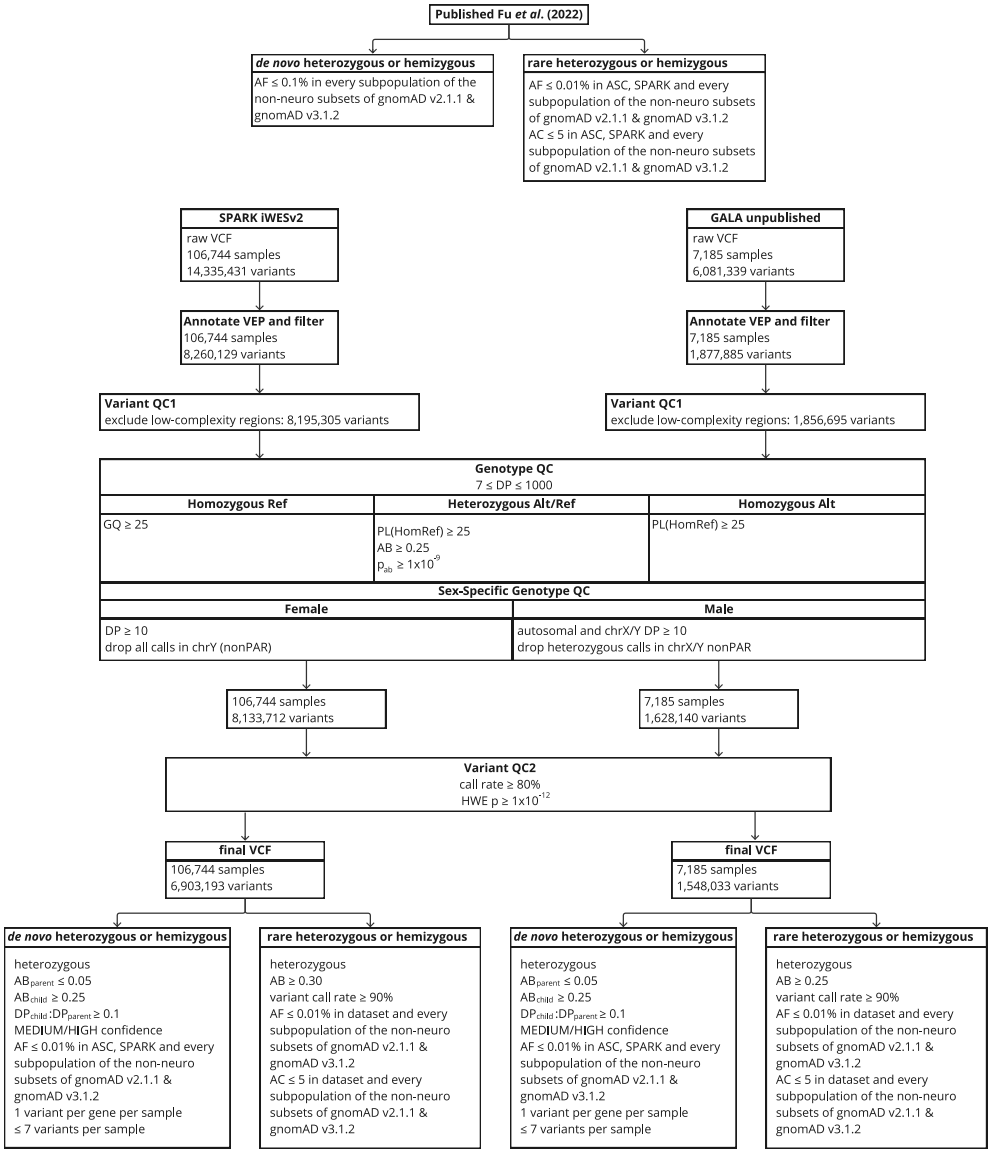
**Extended data** is available for this paper at https://doi.org/10.1038/s41591-026-04228-6.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41591-026-04228-6.

**Correspondence and requests for materials** should be addressed to Joseph D. Buxbaum.

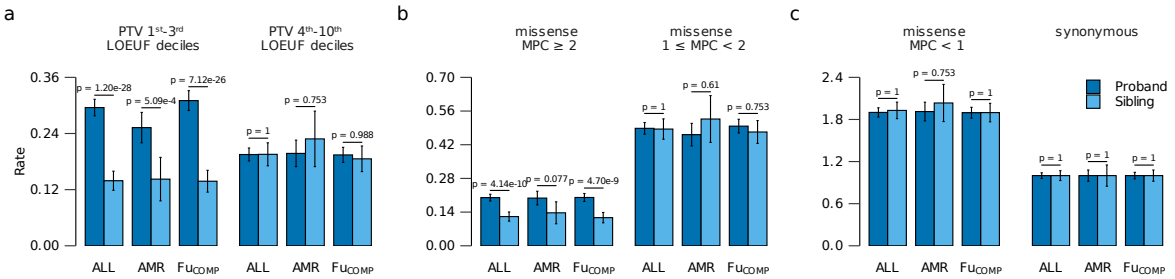
**Peer review information** *Nature Medicine* thanks Andres Moreno-Estrada and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Anna Ranzoni, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-026-04228-6



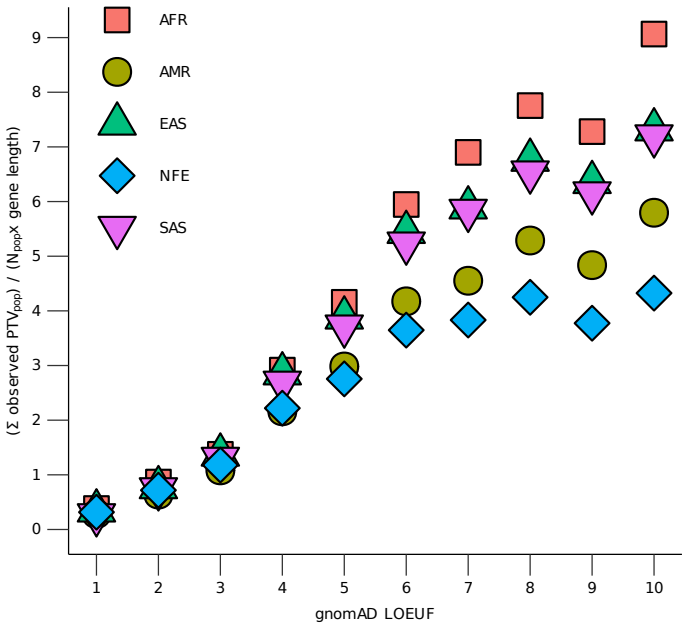

**Extended Data Fig. 1 | Data processing for samples from three different data sources.** The figure describes the variant, genotype, and sample quality control steps that were implemented to process the raw, joint-genotyped VCFs and generate the de novo and inherited calls used for downstream analyses. Sample counts are tabulated before downstream ancestry filtering.

https://doi.org/10.1038/s41591-026-04228-6



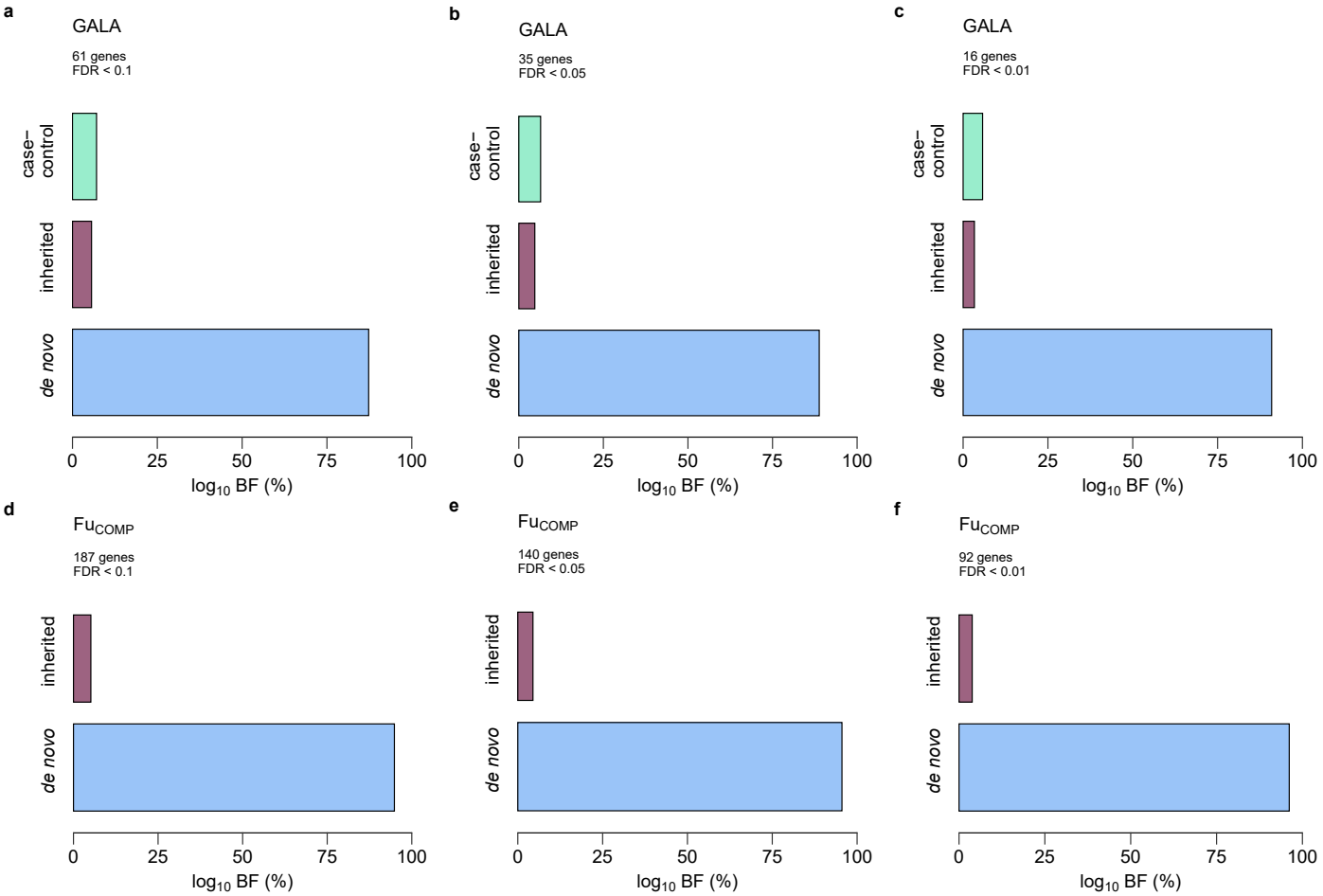

**Extended Data Fig. 2 | Comparison of rare de novo variant counts per sample between ASD probands and unaffected siblings across different ancestries, normalized to synonymous variant rates.** The average number of rare variants per sample –normalized by the synonymous de novo variant rate– is compared between ASD probands and their unaffected siblings for all ancestries (ALL: 17,480 probands and 6,208 siblings), Admixed American (AMR: 4,450 probands and 1,459 siblings), and non-Admixed American ($Fu_{COMP}$: 13,030 probands and 4,749 siblings). The analysis considers: (**a**) protein truncating variants (PTVs) in highly constrained genes (LOEUF deciles 1–3, 5,363 genes) and less constrained genes (LOEUF deciles 4–10, 12,765 genes); (**b**) missense variants categorized by predicted functional severity (MPC ≥ 2 for high severity, 1 ≤ MPC < 2 for moderate severity); and (**c**) MPC < 1 (for low severity) and synonymous missense variants. Data are presented as mean values ± 95% confidence intervals. Statistical significance was assessed using two-sided z-tests comparing normalized de novo mutation rates between probands and siblings. *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) method, and exact adjusted *P* values are shown above the bars.

https://doi.org/10.1038/s41591-026-04228-6



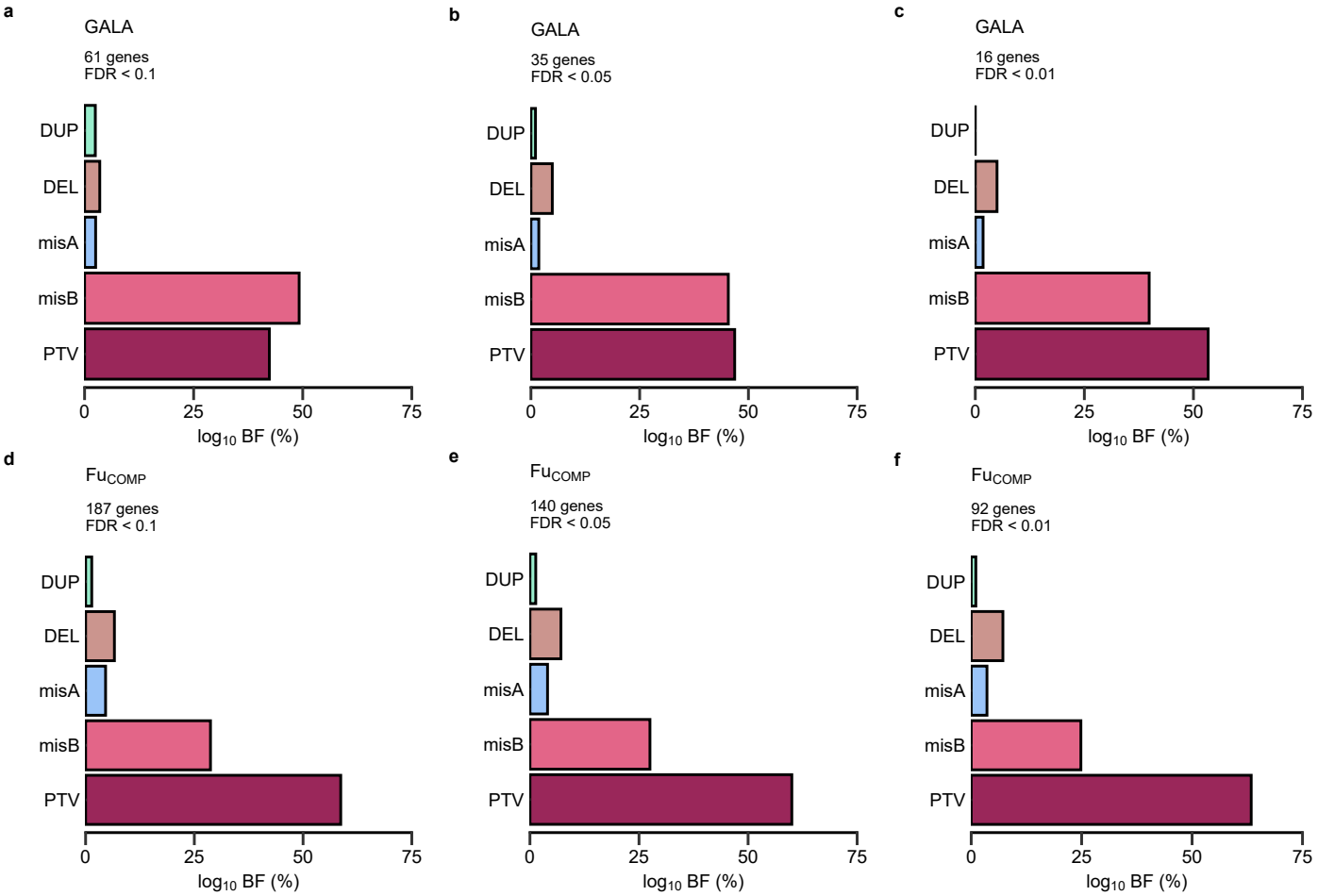

**Extended Data Fig. 3 | Genic burden of PTVs across different ancestries in gnomAD v2.1.1 as a function of gene constraint.** The sum of observed PTVs per ancestry is plotted, scaled to each population's size and total gene coding sequence length within gnomAD LOEUF deciles. The plot includes African American (AFR, nN = 8,128), Admixed American (AMR, nN = 17,296), East Asian (EAS, n = 9,197), Non-Finnish European (NFE, n = 56,885), and South Asian (SAS, n = 15,308) ancestries. LOEUF deciles represent levels of gene constraint, with lower deciles indicating more constrained genes.

https://doi.org/10.1038/s41591-026-04228-6



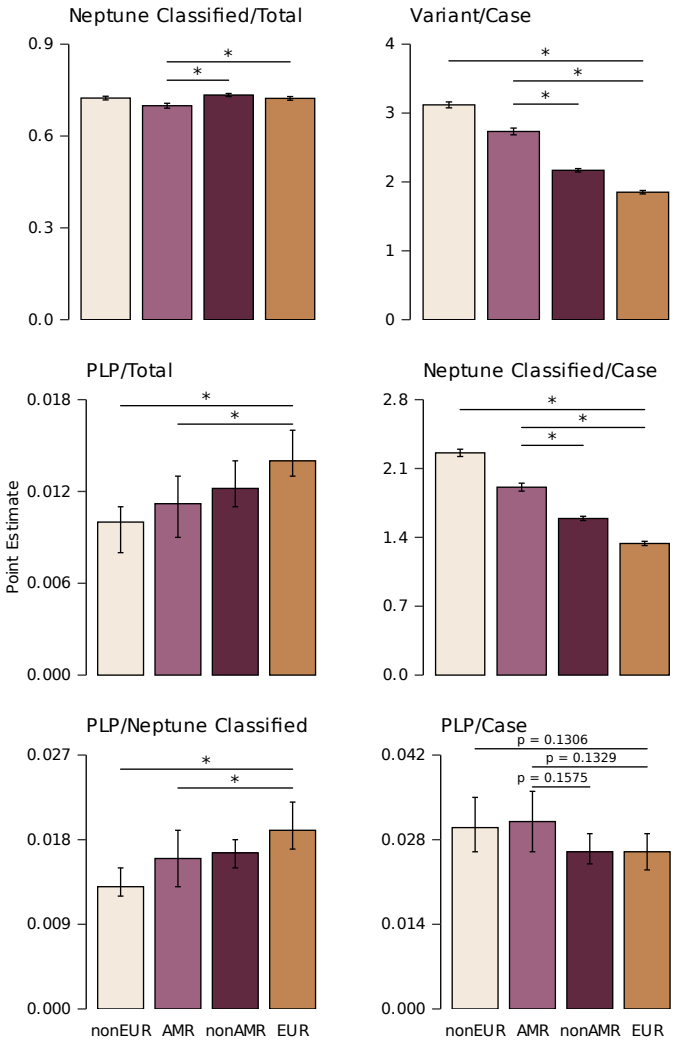

**Extended Data Fig. 4 | Relative contribution to TADA signal by mode of inheritance.** The proportional impact of each inheritance mode on the ASD-associated genes is shown at three false discovery rate (FDR) thresholds: ≤0.1 (**a, d**), ≤0.05 (**b, e**), and ≤0.01 (**c, f**). Panels (**a**–**c**) display results for the GALA cohort, while panels (**d**–**f**) show results for the $Fu_{COMP}$ subset from Fu et al.[4]. BF, Bayes Factor.

https://doi.org/10.1038/s41591-026-04228-6



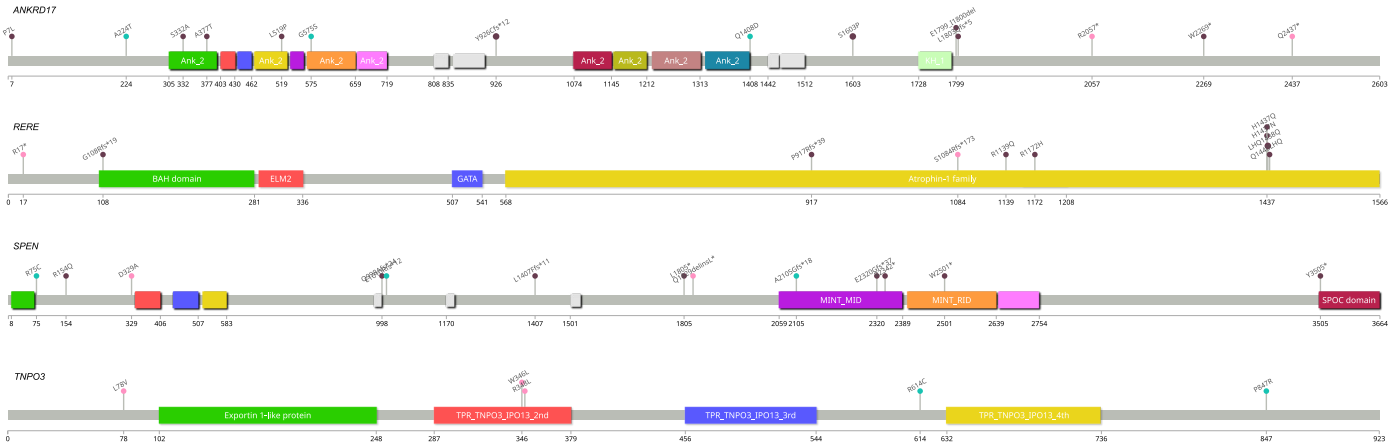

**Extended Data Fig. 5 | Relative contribution to TADA signal by variant type.** The proportional impact of each variant type on the ASD-associated genes is shown at three false discovery rate (FDR) thresholds: ≤0.1 (**a, d**), ≤0.05 (**b, e**), and ≤0.01 (**c, f**). Panels (**a**–**c**) display results for the GALA cohort, while panels (**d**–**f**) show results for the $Fu_{COMP}$ subset from Fu et al.[4]. BF, Bayes Factor.

https://doi.org/10.1038/s41591-026-04228-6



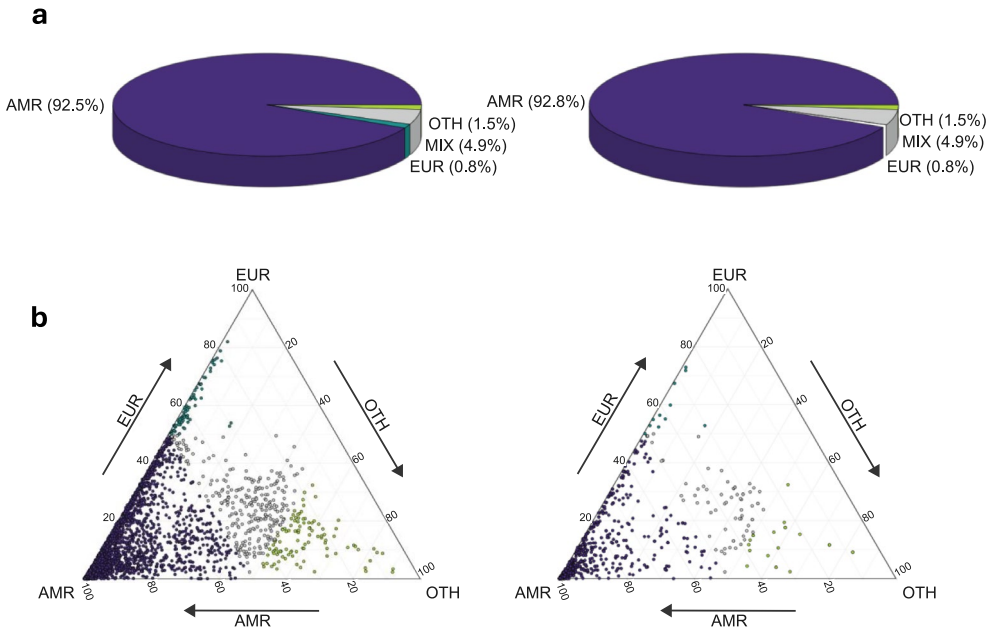

**Extended Data Fig. 6 | Classification rates and proportions of P/LP variants across AMR and non-AMR populations using Neptune.** The figure compares the classification rates and proportions of P/LP variants in the indicated subsamples. Left: The ratio of (**upper**) classified variants (by Neptune) to total variants, (**middle**) P/LP variants to total variants, and (**lower**) P/LP variants to Neptune classified variants is shown for AMR, non-AMR, non-European (non-EUR) and EUR ancestries. Right: Comparisons include (**upper**) the total number of variants, (**middle**) the number of classified variants, and (**lower**) the number of P/LP variants, all expressed per proband. AMR participants have more variants per individual (both total and Neptune-classified) compared to non-AMR participants, but a reduced ability of Neptune to classify variants in AMR contributes to a slightly lower proportion of P/LP variants per individual. Similar results are seen for non-EUR versus EUR. Data are presented as mean values ± 95% confidence intervals (error bars show the plotted CI bounds). Statistical analysis: pairwise two-sided *z*-tests were used to compare groups within each panel; *P* values were adjusted for multiple comparisons using the Benjamini–Hochberg FDR procedure. Asterisks indicate adjusted $P < 0.05$.

https://doi.org/10.1038/s41591-026-04228-6




**Extended Data Fig. 7 | Lollipop diagrams illustrating variants identified in emerging autism-associated genes.** Variants observed in GALA analyses of AMR individuals are marked with pink circles, those found in $Fu_{COMP}$ individuals are marked with green, and variants found in DECIPHER are in purple. Figures were generated using the lollipop software package[52].

https://doi.org/10.1038/s41591-026-04228-6




**Extended Data Fig. 8 | Evaluation of ancestry composition and variant burden among GALA probands.** (**a**) Ancestry proportions for all GALA individuals (**left**) and among carriers of damaging rare variants (**right**) inferred using a Random Forest classifier trained on 1000 Genomes + Human Genome Diversity Project (HGDP) reference populations. Most individuals display majority Admixed American (AMR) ancestry. (**b**) Ternary plots showing the distribution of ancestry proportions among all GALA individuals (**left**) and among carriers of damaging rare variants (**right**).

https://doi.org/10.1038/s41591-026-04228-6

## Extended Data Table 1 | Observed and expected values for 17 genes identified in both GALA and $Fu_{COMP}$

| | | | | PTV | | | | MisB | | | | MisA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | pLI | LOEUF | LOEUF bin | $GALA_{dn}$ | $Fu_{COMPdn}$ | $GALA_{exp}$ | $GALA_{dev}$ | $GALA_{dn}$ | $Fu_{COMPdn}$ | $GALA_{exp}$ | $GALA_{dev}$ | $GALA_{dn}$ | $Fu_{COMPdn}$ | $GALA_{exp}$ | $GALA_{dev}$ |
| *PTEN* | 0.26 | 0.51 | 2 | 4 | 3 | 0.956 | 3.044 | 7 | 9 | 3.339 | 3.661 | 0 | 2 | 0.750 | -0.750 |
| *SHANK3* | 1.00 | 0.12 | 0 | 6 | 12 | 3.822 | 2.178 | 0 | 0 | 0.000 | 0.000 | 0 | 0 | 0.000 | 0.000 |
| *SCN2A* | 1.00 | 0.13 | 0 | 2 | 18 | 5.733 | -3.733 | 6 | 16 | 5.935 | 0.065 | 0 | 3 | 1.125 | -1.125 |
| *CHD8* | 1.00 | 0.08 | 0 | 4 | 19 | 6.052 | -2.052 | 2 | 6 | 2.226 | -0.226 | 1 | 3 | 1.125 | -0.125 |
| *SYNGAP1* | 1.00 | 0.05 | 0 | 4 | 15 | 4.778 | -0.778 | 2 | 4 | 1.484 | 0.516 | 2 | 0 | 0.000 | 2.000 |
| *FOXP1* | 1.00 | 0.18 | 0 | 3 | 10 | 3.185 | -0.185 | 1 | 1 | 0.371 | 0.629 | 1 | 0 | 0.000 | 1.000 |
| *DYNC1H1* | 1.00 | 0.08 | 0 | 4 | 3 | 0.956 | 3.044 | 0 | 11 | 4.081 | -4.081 | 1 | 3 | 1.125 | -0.125 |
| *ADNP* | 1.00 | 0.12 | 0 | 3 | 18 | 5.733 | -2.733 | 0 | 0 | 0.000 | 0.000 | 0 | 0 | 0.000 | 0.000 |
| *TCF4* | 1.00 | 0.22 | 0 | 1 | 2 | 0.637 | 0.363 | 2 | 0 | 0.000 | 2.000 | 0 | 0 | 0.000 | 0.000 |
| *GRIN2B* | 1.00 | 0.06 | 0 | 2 | 7 | 2.230 | -0.230 | 0 | 4 | 1.484 | -1.484 | 2 | 2 | 0.750 | 1.250 |
| *CHD2* | 1.00 | 0.07 | 0 | 2 | 9 | 2.867 | -0.867 | 1 | 5 | 1.855 | -0.855 | 0 | 2 | 0.750 | -0.750 |
| *SATB2* | 1.00 | 0.09 | 0 | 2 | 0 | 0.000 | 2.000 | 0 | 5 | 1.855 | -1.855 | 0 | 1 | 0.375 | -0.375 |
| *TCF20* | 1.00 | 0.10 | 0 | 2 | 4 | 1.274 | 0.726 | 0 | 0 | 0.000 | 0.000 | 0 | 0 | 0.000 | 0.000 |
| *CUL3* | 1.00 | 0.23 | 0 | 2 | 5 | 1.593 | 0.407 | 0 | 0 | 0.000 | 0.000 | 1 | 1 | 0.375 | 0.625 |
| *NRXN1* | 1.00 | 0.25 | 0 | 0 | 4 | 1.274 | -1.274 | 1 | 0 | 0.000 | 1.000 | 0 | 1 | 0.375 | -0.375 |
| *DLG4* | 1.00 | 0.24 | 0 | 2 | 4 | 1.274 | 0.726 | 0 | 1 | 0.371 | -0.371 | 0 | 1 | 0.375 | -0.375 |
| *SCN1A* | 1.00 | 0.07 | 0 | 0 | 2 | 0.637 | -0.637 | 1 | 0 | 0.000 | 1.000 | 1 | 5 | 1.875 | -0.875 |
| *sum* | | | | **43** | **135** | **43.001** | **-0.001** | **23** | **62** | **23.001** | **-0.001** | **9** | **24** | **9.000** | **0.000** |

https://doi.org/10.1038/s41591-026-04228-6

## Extended Data Table 2 | Summary of emerging and notable gene-level findings in GALA

| Gene Symbol | Gene Name | Chromosomal Location | Findings in GALA and Other Cohorts[a] | Associated Disorder | Functional Notes | Comments |
|---|---|---|---|---|---|---|
| *RERE* | arginine-glutamic acid dipeptide repeats | 1p36.23 | Two PTVs in GALA **(Extended Data Figure 7)** | Known ID/ASD gene: Neurodevelopmental disorder with or without anomalies of the brain, eye, or heart, AD (MIM 616975), LoF | Member of the Atrophin family of transcriptional repressors. | *RERE* is located in the proximal 1p36 deletion syndrome critical region; the phenotype of patients with *RERE* variants is reminiscent of that observed in patients with 1p36 deletion syndrome. There is emerging evidence of possible genotype-phenotype correlations in *RERE*, with missense variants in the atrophin-1 domain associated with more severe, syndromic presentations, while PTVs are associated with a less severe phenotype (PMID: 29330883). The two PTVs in GALA participants are consistent with these findings. |
| *MTOR* | mechanistic target of rapamycin kinase | 1p36.22 | Signal in GALA and DD; not in $Fu_{COMP}$ | Known ID/ASD gene: Smith-Kingsmore syndrome, AD (MIM 616638), GoF | MTOR signaling pathway. MTOR signaling is a prominent pathway amongst the top GALA genes, including *PTEN*, *MTOR*, *TSC2*, *MARK2*, and *GLS*. Additionally, CUL3 is a core component of the E3 ubiquitin ligase complex and indirectly regulates MTOR signaling. | Neurodevelopmental disorder associated with ASD caused by gain-of-function missense variants. |
| *SPEN* | spen family transcriptional repressor | 1p36.21-p36.13 | Signal in GALA, trending signal in $Fu_{COMP}$, and strong signal in DD. One *de novo* PTV and three *de novo* missense variants in AMR individuals **(Extended Data Figure 7)** | Known ID/ASD gene: Radio-Tartaglia syndrome, AD (MIM 619312), LoF | Transcriptional repressor; interacts with HDAC1 and HDAC2 and is involved in X-chromosome silencing. | Truncating variants cause Radio-Tartaglia syndrome; the significance of missense variants has not been determined. |
| *GLS* | glutaminase | 2q32.2 | Two *de novo* missense variants (p.Met465Ile, p.Glu596Gln) in AMR individuals | Associated with dominant and recessive NDDs: *de novo* missense variants leading to GLS hyperactivity cause CASGID syndrome, AD (MIM 618339; also known as Infantile cataract, skin abnormalities, glutamate excess, and impaired intellectual development); biallelic loss of function variants cause Global developmental delay, progressive ataxia, and elevated glutamine, AR (MIM 618412), and Developmental and epileptic encephalopathy 71, AR (MIM 618328). | Encodes glutaminase, the enzyme converting glutamine into glutamate; regulates MTOR signaling. The presence of *GLS*, *GRIN1*, *GRIN2B*, *SYNGAP1*, *DLG4*, *SHANK3*, *NRXN1*, *SCN2A*, *SCN1A*, and *CACNA1D* among the top genes in our analysis is consistent with a large body of evidence implicating glutamatergic synaptic organization, plasticity and signaling in ASD. | A single *de novo* heterozygous *GLS* variant leading to enhanced catalytic activity has been reported (PMID: 30239721). Studying glutamine and glutamate levels in the urine of the two participants with *de novo* missense variants in *GLS*, or using brain magnetic resonance spectroscopy, would clarify the significance of these variants. |
| *CACNA1D* | calcium voltage-gated channel subunit alpha1 D | 3p21.1 | Two *de novo* missense variants in AMR individuals: p.Tyr1062Cys and p.Gly1149Val | Primary aldosteronism, seizures, and neurologic abnormalities, AD (MIM 615474), GoF | Voltage-gated L-type Cav1.3 calcium channel | The two *de novo* *CACNA1D* missense variants in our AMR dataset (p.Tyr1062Cys and p.Gly1149Val) are deemed pathogenic using a recently described *CACNA1D*-specific model (PMID: 38553610). |
| *PAK2* | p21 (RAC1) activated kinase 2 | 3q29 | *De novo* frameshift (p.Pro30Leufs*8) and missense (p.Met436Lys) in AMR individuals | Provisional disease association: Knobloch syndrome 2, AD (MIM 605022), based on a report of two brothers with a *de novo* missense variant (p.Glu435Lys) and severe ocular defects, one of which also had ID and ASD (PMID: 33693784). | Serine/threonine kinase that activates the LIMK-cofilin pathway, a key regulator of synaptic actin dynamics (PMID: 33306168). | The *de novo* missense variant (p.Met436Lys) identified in an ASD individual is located in the protein kinase domain, adjacent to the variant previously reported in Knobloch syndrome 2 (p.Glu435Lys), which leads to substantial loss of kinase activity (PMID: 33693784). Our results, if validated with functional analysis, would provide further evidence that *PAK2* could be an ASD-associated gene. |
| *ANKRD17* | ankyrin repeat domain 17 | 4q13.3 | Two *de novo* PTVs and one missense variant in AMR individuals **(Extended Data Figure 7)** | Known NDD gene: Chopra-Amiel-Gordon syndrome, AD (MIM 619504), LoF | Ankyrin repeat domain protein | ASD has been observed in some cases of Chopra-Amiel-Gordon syndrome, and our results in GALA and in $Fu_{COMP}$ support this gene as an ASD-associated gene. |
| *YWHAG* | tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein gamma | 7q11.23 | Significant in GALA and DD; not in $Fu_{COMP}$. Two *de novo* missense variants in AMR individuals: p.Arg57Cys (recurrent pathogenic variant; also present in a patient in $Fu_{COMP}$ and in DECIPHER) and a nearby variant, p.Lys50Gln (uncertain significance) **(Figure 5)** | Developmental and epileptic encephalopathy 56, AD (MIM 617665), GoF | The 14-3-3 family of proteins are scaffolding proteins that integrate signals in neurons and other cells (PMID: 38685162); *YWHAG* encodes 14-3-3γ. | The p.Arg57Cys variant identified in an AMR individual has been reported as pathogenic in 8 individuals in ClinVar. |
| *TNPO3* | transportin 3 | 7q32.1 | Across GALA and FuCOMP there were five de novo missense variants (Extended Data Figure 7) | Significant in GALA and DD; not in $Fu_{COMP}$. Two *de novo* missense variants in AMR individuals: p.Arg57Cys (recurrent pathogenic variant; also present in a patient in $Fu_{COMP}$ and in DECIPHER) and a nearby variant, p.Lys50Gln (uncertain significance) **(Figure 5)** | Transports target proteins into the nucleus, including splicing factors. | The five *de novo* missense variants identified in ASD individuals are distributed throughout the protein, in contrast to the frameshift variants that extend the C-terminus described in limb-girdle muscular dystrophy. While mutations in *DMD* (Duchenne and Becker muscular dystrophy) can lead to ASD (PMID: 38701157), further genotype–phenotype studies are needed to understand whether different mutations in *TNPO3* lead to psychiatric and/or muscular phenotypes. |
| *MARK2* | microtubule affinity regulating kinase 2 | 11q13.1 | FDR < 0.002 in GALA; not significant in $Fu_{COMP}$; signal in DD **(Figure 5)** | Recently (since submission of this manuscript) identified ID/ASD gene: Intellectual developmental disorder, autosomal dominant 76 (MIM 621285), LoF | Encodes a serine/threonine kinase that plays a key role in phosphorylation of microtubule associated proteins and regulates the MTOR pathway (PMID: 38302729). | Results from the GALA cohort provide genome-wide support for *MARK2* as an ASD-associated gene. |
| *PACS1* | phosphofurin acidic cluster sorting protein 1 | 11q13.1-q13.2 | Signal in GALA and DD; not in $Fu_{COMP}$. Two *de novo* missense variants in GALA: p.Arg203Trp (recurrent pathogenic variant) and p.Arg709Gln (uncertain significance) **(Figure 5)** | Known ID/ASD gene: Schuurs–Hoeijmakers syndrome, AD (MIM 615009), GoF | *PACS1* variants have a direct impact on neuronal activity by increasing the interaction between PACS1 and HDAC6 and potentiating its deacetylase activity (PMID: 37848409). | *PACS1* neurodevelopmental disorder is caused by a recurrent *de novo* missense variant (p.Arg203Trp) exerting a GoF effect. DECIPHER reports multiple *de novo* p.Arg203Trp variants, but also identifies other *de novo* changes. |
| *DYNC1H1* | dynein cytoplasmic 1 heavy chain 1 | 14q32.31 | Four *de novo* PTVs and one missense variant in AMR individuals | Known NDD gene: Cortical dysplasia, complex, with other brain malformations 13, AD (MIM 614563). Also involved in two distal motor neuronopathies: Charcot-Marie-Tooth disease, axonal, type 2O, AD (MIM 614228), and Spinal muscular atrophy, lower extremity-predominant 1, AD (MIM 158600) | Encodes dynein heavy chain 1, which forms the core of cytoplasmic dynein – the main retrograde motor in cells (PMID: 33705456). | Dominant pathogenic variants in *DYNC1H1*, mostly missense, cause early and later onset neurodevelopmental and peripheral neuromuscular disorders (PMID: 38848546). |
| *GSE1* | Gse1 coiled-coil protein | 16q24.1 | FDR = 0.0005 in GALA; three *de novo* PTVs in AMR individuals **(Figure 5)** | Not previously linked to human disease | Forms a complex with the HDAC1/CoREST deacetylase/demethylase co-repressor complex (PMID: 37878419). | *GSE1* has not been associated with any human diseases, but an examination of the newest (and largest) ASC dataset shows genome-wide significance (FDR=0.00091) for this gene, with five *de novo* PTVs in cases and none in unaffected siblings (ASC, in preparation). *GSE1* is significantly constrained for PTVs (pLI = 1, LOEUF 0.36). Provides additional support for gene expression regulation in ASD. |

[a]Findings in Latin American ASD individuals in GALA, non-Latin American ASD individuals in Fu et al.[4] ($Fu_{COMP}$), and in a large cohort of developmental disorders[29]. Abbreviations: AD, autosomal dominant; AMR, admixed American ancestry; AR, autosomal recessive; ASC, Autism Sequencing Consortium; ASD, autism spectrum disorder; DD, developmental disorder; FDR, false discovery rate; GoF, gain of function; ID, intellectual disability; LoF, loss of function; NDD, neurodevelopmental disorder; PTVs, protein truncating variants.

nature portfolio

Corresponding author(s): Joseph D Buxbaum

Last updated by author(s): Nov 21, 2025

# Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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<br>*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<br>*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 | N/A |
| Data analysis | All software used in this study is publicly available at the cited references. The R code used to generate the TADA analysis and figures is available under the MIT license at https://github.com/buxbaum-lab/GALA |

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequencing data for the ASC and GALA samples are available through controlled access via dbGaP (accession phs002502) and on the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) under Study Accession phs002502.v1.p1 (https://anvilproject.org/data). SPARK phenotype and sequencing data are available to authorized users through the SFARI Base (https://www.sfari.org/resource/sfari-base/).

Individual-level data from the ASC and GALA cohorts are not publicly available due to participant privacy restrictions. Researchers may request access by contacting Dr. Joseph D. Buxbaum (joseph.buxbaum@mssm.edu). All requests will be reviewed by the Mount Sinai Institutional Data Access Committee to ensure compliance with participant consent and IRB protocols. Reasonable requests will receive a response within two to four weeks. Summary variant counts, gene-level burden statistics, and figure source data are available in the accompanying supplementary materials and at https://github.com/buxbaum-lab/GALA.

## Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

| | |
|---|---|
| Reporting on sex and gender | Included. Sex was recorded based on self-report at enrollment and confirmed with genetic information. Both male and female participants were included; however, sex-stratified analyses were not conducted, as the primary focus of this study was on de novo and rare variant burden across ancestry groups rather than sex differences. Participant ages varied by cohort, with probands typically enrolled during childhood or adolescence and parents as adults. |
| Reporting on race, ethnicity, or other socially relevant groupings | No race or ethnicity was collected for the study. |
| Population characteristics | Included. |
| Recruitment | Included. |
| Ethics oversight | This study was conducted in accordance with Nature Portfolio's guidelines on inclusion and ethics in global research. The research was designed to include participants of diverse ancestries, with the goal of improving representation in autism genetics research. Study protocols were approved by the Institutional Review Boards (IRBs) at all participating sites, including the Program for the Protection of Human Subjects at the Icahn School of Medicine at Mount Sinai (GCO# 14-1082(0001)), as well as the local IRBs in Brazil, Colombia, Peru, Mexico, Kaiser Permanente (KP), and the CHARGE study (see above, Cohort Description). Written informed consent was obtained from all participants or from parents or legal guardians of minors. Data collection adhered to relevant ethical and cultural standards and compensation for participation varied by site. Collaborations between institutions in the United States and Latin America were established to ensure equitable contributions across sites. Local investigators in Brazil, Colombia, Mexico, and Peru were involved in data collection and authorship. |

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

# 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 | 15427 ASD samples were included in the current analysis. All available samples of Latin American ancestry were used. |
| Data exclusions | Samples were excluded if the probands were not of Latin American ancestry, if the sex mismatched between reported and inferred, or if the sequencing data failed quality control. |
| Replication | All analyses in this study were designed to be fully reproducible, and we took several measures to verify the robustness of the findings. All computational workflows were executed using version-controlled pipelines with fixed software versions, documented parameters, and publicly available code. Key analytical steps—including variant calling, quality control, association testing, and downstream statistical analyses—were rerun independently to confirm that results were consistent. |
| Randomization | Observational study. |
| Blinding | Observational study. |

# 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 and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |
| ☒ | ☐ Plants |

## Methods

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

# Plants

Seed stocks

*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*

Novel plant genotypes

*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*

Authentication

*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*