

CNV-GWAS identifies 14 novel Autism-associated loci and distinct effects on Cognitive Ability

Cecile Poulain

`cecile.poulain@umontreal.ca`

Centre de Recherche du CHU Sainte-Justine

Thomas Renne

Centre de Recherche du CHU Sainte-Justine

Vincent Bourque

Centre de Recherche du CHU Sainte-Justine

Jade England

Centre de Recherche du CHU Sainte-Justine

Florian Benitiere

Centre de Recherche du CHU Sainte-Justine

Audrey Lemaçon

Centre de Recherche du CHU Sainte-Justine

Catherine Proulx

Centre de Recherche du CHU Sainte-Justine

Zohra Saci

Centre de Recherche du CHU Sainte-Justine

Omar Shanta

University of California San Diego

Worrawat Engchuan

The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

Simon R. Cox

Lothian Birth Cohorts, Department of Psychology, University of Edinburgh

David Porteous

Lothian Birth Cohorts, Department of Psychology, University of Edinburgh

Gail Davies

Lothian Birth Cohorts, Department of Psychology, University of Edinburgh

Paul Redmond

Lothian Birth Cohorts, Department of Psychology, University of Edinburgh

Sarah E. Harris

Lothian Birth Cohorts, Department of Psychology, University of Edinburgh

Gunter Schumann

King's College London

Zdenka Pausova

Hospital for Sick Children

Tomas Paus

Centre de Recherche du CHU Sainte-Justine

Celia Greenwood

McGill University

Laura Almasy

Children's Hospital of Philadelphia

David Glahn

Yale University

Guillaume Huguet

Centre de Recherche du CHU Sainte-Justine

Sebastien Jacquemont

Centre de Recherche du CHU Sainte-Justine

Research Article

Keywords: CNV, GWAS, ASD, Cognitive ability

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10933884/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

CNV-GWAS identifies 14 novel Autism-associated loci and distinct effects on Cognitive Ability

AUTHORS:

Cécile Poulain^{1,2}; Thomas Renne^{1,2}; Vincent-Raphaël Bourque^{1,3}; Jade England¹; Florian Bénitière¹; Audrey Lemaçon¹; Catherine Proulx^{1,2}; Zohra Saci¹; Omar Shanta⁴; Worrawat Engchuan⁵; Simon R. Cox⁶; David Porteous^{6,7,8}; Gail Davies⁶; Paul Redmond⁶; Sarah E. Harris⁶; Gunter Schumann⁹; Zdenka Pausova^{10,11,12}; Tomas Paus^{1,13}; Celia M T Greenwood^{14,15,16}; Laura Almasy¹⁷; David C Glahn¹⁸; Guillaume Huguet^{**1}; Sébastien Jacquemont^{**1,19}

1 Centre Hospitalier Universitaire Sainte-Justine Research Center, Montreal, Quebec, Canada

2 Bioinformatics Graduate Program, Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, Quebec, Canada

3. Department of Psychiatry, McGill University, Montreal, Quebec, Canada

4 Department of Psychiatry, University of California San Diego, La Jolla, CA, USA

5 The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

6 Lothian Birth Cohorts, Department of Psychology, School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, EH8 9JZ, UK.

7 Medical Genetics Section, Centre for Genomic & Experimental Medicine, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh, EH4 2XU, UK.

8 Generation Scotland, Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK.

9 Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, England

10 Research Institute of the Hospital for Sick Children, Toronto, ON, Canada.

11 Departments of Physiology and Nutritional Sciences, University of Toronto, Toronto, ON, Canada.

12 ECOGENE-21, Chicoutimi, Quebec, Canada.

13 Department of Psychiatry and Addictology, Department of Neuroscience, Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada

14 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

15 Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, QC, Canada

16. Gerald Bronfman Department of Oncology, McGill University, Montreal, Quebec, Canada

17 Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

18 Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA; Olin Neuropsychiatric Research Center, Institute of Living, Hartford Hospital, Hartford, CT, USA.

19 Department of Pediatrics, Université de Montréal, Montreal, Quebec, Canada

Corresponding author:

Cécile Poulain, cecile.poulain@umontreal.ca

Sainte-Justine University Hospital,

3175 chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 1C5, Canada

Sébastien Jacquemont, sebastien.jacquemont@umontreal.ca

Sainte-Justine University Hospital,

3175 chemin de la Côte-Sainte-Catherine, Montréal, QC H3T 1C5

Abstract (234 words)

Keywords: CNV, GWAS, ASD, Cognitive ability.

Copy number variants (CNVs) are well-established risk factors for Autism Spectrum Disorder (ASD). Prior rare-variant association studies have predominantly relied on family-based cohorts, limiting statistical power and identification of novel loci. In addition, it has been challenging to distinguish the effects of CNVs for ASD from the impact on cognitive ability. Here we present a large CNV association study of ASD, analyzing 29,696 autistic and 528,576 control individuals, with cognitive ability data available for over half of the participants. We used two complementary designs, leveraging related and unrelated controls. The unrelated case-control analysis yielded substantially more associations than the family-based design, while effect estimates remained concordant across approaches and showed no evidence of stratification bias. Among recurrent neurodevelopmental CNVs, we identified 19 ASD associations, 8 of which had not been previously linked to ASD, including duplications at 7q11 (*ZNF92*) and 9p24 (*DMRT1*), previously associated with schizophrenia. Across the genome, we identified 14 novel candidate CNV loci, several previously associated with risk for schizophrenia, ADHD, or other psychiatric conditions by common and rare variant studies. Cognitive ability data allowed us to delineate two distinct genetic architectures: CNVs predominantly associated with ASD and other mental health conditions, which were enriched for duplications and encompassed less-constrained genes; and CNVs affecting both ASD and cognition, which were enriched for deletions. Together, these findings suggest that classes of variants, rather than genes, may show preferential phenotypic associations.

INTRODUCTION (531 words)

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by atypical social communication and interaction, intense interests or repetitive behaviors¹.

Rare genetic variants are a major contributor to ASD risk²⁻⁴. Among these, Copy Number Variants (CNVs) (deletions or duplications > 1000 base pairs) are routinely screened in neurodevelopmental clinics and identified in 5% to 11% of autistic individuals⁵. To date, case-control studies have been performed in family-based cohorts of up to 13,000 autistic individuals^{3,6}, formally identifying 16 CNVs across 13 genomic loci⁷ associated with ASD. While the increased CNV burden in autism^{4,8-11} suggests that many more loci are linked to the condition, novel associations have been challenging to identify^{3,4,8,12}.

Intellectual disability (ID) is one of the most prevalent co-occurring conditions in children with ASD¹³. Many, if not all ASD-associated CNVs modulate cognitive ability in both autistic individuals⁷ and the general population^{14,15}. As a result, it has been difficult to disentangle the respective contributions of CNVs to autism and cognitive ability¹⁶. Previous studies have suggested that some genes may be preferentially associated with ASD when affected by deleterious variants, whereas other genes may be primarily associated with cognitive impairment^{3,7}, but such distinction remains subject to debate¹⁶. In particular, ASD genetic association studies have not included assessments of cognitive ability in controls, thereby preventing adjustment of genetic associations for cognitive ability.

In contrast to other psychiatric conditions studied through population-based designs (i.e., controls are sampled from the population), large ASD genetic association studies were mostly conducted using family-based cohorts^{3,6,17}. Family designs have the advantage of enabling the detection of *de novo* variants as well as providing robust control for stratification biases and indirect parental genetic effects^{18,19}. However, family-based designs usually provide less statistical power than population-based studies, as assembling large family cohorts is expensive and logistically difficult^{18,19}. In addition, family-based designs are restricted to relatives, which reduces the genetic variance captured relative to population-based studies (i.e., the effects of risk variants transmitted by parents may be underestimated).

In this study, we aimed to increase power to detect novel associations by leveraging the largest ASD family-based sample available to date, as well as controls from unselected population

cohorts. To distinguish preferential effects on ASD risk versus cognition, we used cognitive ability assessments in controls.

We therefore conducted two complementary analyses in parallel: a family-based analysis comparing autistic individuals with unaffected relatives (N = 68,605; ASD = 25,888; Controls = 42,717), and an unrelated case-control analysis comparing autistic probands with individuals from unselected population cohorts (N = 512,690; ASD = 29,696; Controls = 482,994).

First, we systematically assessed the association between recurrent CNVs and autism across our two complementary study designs, with the unrelated case-control analysis identifying 19 ASD-associated CNVs, including eight that had not previously been associated with ASD. We also identified that some recurrent CNVs are preferentially associated with ASD or cognitive ability. We then conducted a genome-wide CNV association study (CNV-GWAS), identifying 14 novel candidate loci associated with autism. Finally, we performed CNV-GWAS analyses across multiple traits, including developmental traits assessed in the autistic population and to phenotypes in the general population to characterize the pleiotropy of ASD-associated CNVs.

RESULTS (1901 words)

Associations between ASD and neurodevelopmental recurrent CNVs.

Studies of recurrent CNVs in ASD have typically been conducted with modest sample sizes across distinct cohorts, and comprehensive large-scale analyses are lacking.

To address this gap, we assembled a large dataset including three autism cohorts and six unselected population cohorts, leveraging both related and unrelated controls (Table 1). To reduce the risk of spurious associations due to array-platform differences, we applied a stringent CNV harmonization procedure, retaining CNVs detected by at least 10 probes across all platforms and excluding those with heterogeneous frequencies across platforms (see Methods). We compiled a comprehensive list of recurrent CNVs across 59 genomic regions based on the ClinGen dosage sensitivity map, with published evidence of clinical relevance²⁰, many of which have previously been associated with neurodevelopmental and psychiatric conditions, including autism^{9,12,21–23} (Table S1). Of these, we assessed autism liability for 56 recurrent CNVs with at least 15 carriers in our dataset by conducting two parallel analyses: one using related controls (N = 68,605; ASD = 25,888; Controls = 42,717) and one using unrelated controls (N = 512,690; ASD = 29,696; Controls = 482,994).

ASD-liability estimates were highly concordant between the two analyses, suggesting limited bias from the inclusion of unrelated controls (ICC = 0.87 [0.74–0.93]; P = 2.10e-08; Figure 1A). While no systematic bias was observed between estimates from both approaches (Figure S1), the unrelated control analysis revealed 5 times more ASD associations (Table 2).

All 4 CNVs associated with ASD using related controls were also identified in the unrelated control analysis, which yielded an additional 15 CNVs (Table 2; Table S2). Eight of these recurrent CNVs were not previously associated with ASD^{3,4,12,17,24–26} in case-control studies (Table 2), including duplications of 7q11 (*ZNF92*), 9p24 (*DMRT1*), and 17q12 (*HNF1B*), which were all previously associated with schizophrenia^{9,27}. Even CNVs with well-established risk, such as 1q21.1 TARs, 3q29 (*DLG1*), and 15q13 BP4-BP5, reached significance only in the unrelated control analysis, highlighting the greater power of the unrelated case control design, likely due to its larger sample size and greater genetic variance. Our results also question previously established links between CNVs and autism. Four CNVs (one deletion: 1q21.1 distal; and three duplications: 7q11.23 Williams Beuren, 17p11.2, and 22q11.2 proximal)²⁶ did not demonstrate any association in our analyses, highlighting the resilience of many carriers identified in unselected cohorts (i.e., 15 unrelated controls with the Williams-

Beuren duplication) known to be heavily biased against neurodevelopmental disorders (Table 2).

We compared our estimates provided by the unrelated control analysis with the population-based estimates from the iPSYCH study²⁸. Despite minor differences in CNV definitions and effect size metrics (odds ratios for the current study vs hazard ratios for iPSYCH), estimates were well correlated (deletions: $r = 0.69$; duplications: $r = 0.63$; Figure 1B). We performed a sensitivity analysis using whole-exome CNV calls from autistic individuals and families in a larger sample of the SPARK cohort ($N = 124,256$; ASD = 47,229; Controls = 77,027; Figure S2), replicating three of the four signals identified in the primary related control analysis (Table S3). Three additional CNVs reaching significance in this CNV exome dataset (15q11.2 deletion and duplication, and the 15q13 BP4-BP5 deletion) were also identified in the unrelated control analysis. We evaluated whether associations between recurrent CNV and ASD were independent of cognitive ability by including cognition as a covariate in the models. Eight of the nineteen associated CNVs remained significant after adjustment (Table 2). We further examined the potential preferential effects of recurrent CNVs on ASD versus cognitive ability. To do so, we computed and compared the effect sizes of CNVs on both traits. While both effects were correlated ($r = -0.69$, $P = 2.87 \times 10^{-8}$; Figure 1C), we could distinguish CNVs with ASD-only association ($N = 5$), cognitive ability-only association ($N = 6$), and those associated with both traits ($N = 10$). We also observed mirror associations at the 1q21.1 distal (ASD associated with duplications and cognitive ability with deletions) and at the 1q21.1 TAR locus (ASD and cognitive ability associated with deletions and duplications, respectively). The 9p24 (*DMRT1*) duplication showed an association with ASD in the absence of any signal for cognitive ability.

Genome-Wide Analysis of CNVs in Autism and Cognitive Ability.

Beyond known recurrent CNVs, we conducted a genome-wide association study (CNV-GWAS) focusing on gene dosage, using CNVs fully encompassing one or more genes.

We tested gene associations using the two analytical designs presented above (using related and unrelated controls), covering up to 1,798 unique genes covered by at least 15 deletions or duplications (Figure 2A). In the unrelated case-control analysis, 494 distinct genes showed significant associations with autism ($FDR\ q < 0.05$) when deleted ($n = 238$) or duplicated ($n = 374$) (Figure 2B). Similarly to the recurrent CNV analysis, the number of associated genes was six times higher in the unrelated compared to the related control analysis. To account for multi-

genic CNVs, genes were grouped into clusters based on overlapping carrier counts (Figure 2A). At the cluster level, we identified a total of 146 significant clusters in the unrelated control analysis (Deletions = 47; Duplications = 99). Of these, 73 did not include known recurrent CNV or SFARI genes (Figure 2C), and all 15 clusters identified in the family-based analysis were replicated using unrelated controls. Down-sampling the unrelated dataset to match the sample size and case:control ratio of the related analysis still yielded 2.4-fold more significant associations (empirical $p < 0.002$ across 500 iterations, Figure S3), demonstrating that differences between designs are not only attributable to sample size. Estimates from the two analytical designs were correlated ($R = 0.45$, $P = 4.43\text{e-}07$; Figure 2D), with a significant bias (mean difference = 0.55, $P = 1.12\text{e-}08$; Figure S4) towards higher estimates in the unrelated case-control analysis. This systematic bias was unchanged when restricting the unrelated analysis to individuals of European ancestry (Figure S4). Sensitivity analysis using whole-exome CNV calls in the larger SPARK sample replicated 10 of the 15 significant gene clusters from the primary family-based analysis (Figure S4, Table S5). In addition, 15 signals only previously identified in the unrelated control analysis reached significance in the larger familial CNV-exome sample (Figure S5). Concordance remained similar between ASD-liability estimates from the larger exome CNV sample and the unrelated control design ($ICC = 0.55$ [0.42-0.65]; $P = 6.73\text{e-}10$, Figure S4). Overall, 30 clusters from the unrelated case-control analysis were replicated in either family-based analysis. Among these, 80 % (24/30) were known rCNV or SFARI genes (any category).

Using individuals from the unrelated control dataset with available cognitive data ($N = 269,227$; ASD = 11,691), we tested whether previously identified ASD associations persisted after adjustment for cognitive ability in the models. Estimates remained highly concordant before and after adjustment ($R = 0.92$; $ICC = 0.94$; Figure S6), and 49 clusters (~33 %; Deletions = 14; Duplications = 35) remained significant after adjustment for cognitive ability (Figure S5). To further investigate this question, we conducted a CNV-GWAS on cognitive ability using population-based controls and individuals with autism. Both effects were negatively correlated ($R = -0.38$, $P = 4.01\text{e-}04$; Figure 2E), with gene clusters impacting autism liability also affecting cognition. We then divided the clusters into three groups: those associated only with ASD, only with cognitive ability, or with both traits. The ASD-only group was enriched in duplications (FDR-adjusted Fisher exact, $OR = 9.9$ [2.88–40.95]; $q = 9.46\text{e-}05$, Table S6), while clusters associated with both traits were enriched in deletions (FDR-

adjusted Fisher exact, OR = 0.20 [0.06–0.64]; $q = 4.34\text{e-}03$, Table S6). ASD-only associated genes (mainly encompassed in duplications) showed lower genetic constraint (measured by LOEUF) compared to the genome-wide distribution. (FDR-adjusted, $p = 2.13\text{e-}09$, Figure 2F, Table S7).

Novel associations with Autism.

To define robust associations, we used a Bonferroni-corrected threshold based on the number of tested clusters (Bonferroni $p < 9.4\text{e-}05$; Figure 2A), identifying 67 significant loci (29 deletion and 38 duplication clusters) in the unrelated case-control analysis. Only 2 of these clusters survived this multiple-testing threshold in the family-based analysis.

Fourteen novel ASD-associated loci were identified after excluding regions with CNVs overlapping known recurrent CNVs or high-confidence SFARI genes (categories 1 and Syndromic), as well as regions that were not replicated when removing non-European participants (Table 3, Figure 3, Figure S7).

We computed the frequencies of *de novo* CNVs across these 14 loci in two large, independent clinical datasets (DECIPHER²⁹ and the CHUSJ diagnostic laboratory) to evaluate their levels of genetic constraint. Significant enrichment of *de novo* events relative to the null model (see Methods) was observed in 3 of the 14 loci, exclusively for deletions, with 2 loci showing a *de novo* proportion exceeding 40%. Of note, for most regions (except 2), genes were broadly expressed across tissues (GTEx, Figure S8), with some regions (e.g., 8p23, 16q23) showing preferential brain expression.

Four of the 14 identified loci overlapped with regions and/or genes with previous suggestive evidence for association with ASD or neurodevelopmental disorders. These included the 8p23 deletion, which spans 6 genes, including 3 SFARI category 2 genes (*ARHGEF10*, *CLN8*, and *DLGAP2*), two of which (*ARHGEF10* and *DLGAP2*) were identified in a few ASD genetic studies^{30,31}. This region was also reported in a 2014 CNV study of developmental delay²³, although none of the genes have been robustly linked by ClinGen to neurodevelopmental or psychiatric disorders. In addition, burden analyses of CNVs encompassing synaptic genes, including *DLGAP2*, have been reported as enriched among individuals with schizophrenia⁹. Similarly, 1q44 duplication contains 38 genes, including 3 SFARI category 2 genes (*OR2M4*, *OR1C1*, *OR2T10*) with limited suggestive evidence from a single CNVs study³² reported for *OR2T10*. Duplications partially overlapping this region showed trends of association with

developmental delay²³. The 8q11 duplication, including *RB1CC1*, was previously associated with Schizophrenia³³, and this gene was also identified as genome-wide significant for schizophrenia in a recent study³⁴. A weak trend toward an association with developmental delay²³ was also reported for the 19p12 duplication encompassing *ZNF675*.

The 11 other loci showed no prior evidence of association between gene-disrupting variants and psychiatric disorders. Among those, 5 exhibited previous associations through GWAS of common variants. A full description of these 14 regions is provided in Table S8.

Pleiotropic effects of CNVs across developmental and population-wide traits

It is well documented that many pathogenic CNVs are associated with multiple diagnoses and traits^{35,36}. We therefore investigated the extent to which ASD-associated gene clusters identified above exhibited pleiotropic effects. We used 45 well-studied heritable phenotypes from the UK Biobank, along with 16 developmental phenotypes and autism-related traits (Table S9). We conducted CNV-GWAS for each trait (Figure 4A). We identified five or more significant associations, either deletions or duplications, for 27 traits across all phenotypes (FDR $q < 0.05$). To assess the level of pleiotropy, we correlated GWAS estimates (the median of gene-level estimates for each independent cluster) across traits, separately for deletions and duplications. (Figure 4B, Table S10). We observed higher correlations (i.e., pleiotropy) for deletions than for duplications. Mental health traits (such as anxiety, neuroticism, loneliness, and miserableness) showed mild to moderate correlation with ASD estimates. ASD-associated clusters also affected traits beyond cognitive and psychopathology domains, including FVC, Forced Expiratory Volume in 1 second (FEV1), standing height and BMI. Among all developmental, cognitive, and behavioural traits assessed in the autistic cohorts, cognitive ability and motor milestones were the only traits showing a correlation with autism.

DISCUSSION (886 word)

By combining study designs, we provide a comprehensive characterization of CNV contributions to ASD liability. The use of unrelated controls substantially increased power, enabling the identification of 8 novel recurrent CNV associations and 14 novel genome-wide loci, several of which had previously been linked to other psychiatric conditions. Beyond ASD liability, comparing CNV effects on autism and cognitive ability revealed preferential associations, suggesting partially distinct genetic architectures underlying these two traits. Finally, ASD-associated CNVs showed broad pleiotropic effects across a wide range of complex traits, extending their impact well beyond the neurodevelopmental domain.

Unrelated controls increase power without introducing substantial bias

We have systematically assessed the contribution of CNVs to ASD using both related and unrelated controls, in the largest sample assembled to date. Most CNV association studies for ASD were conducted using a family-based design^{6,17,31,37,38}, which provides robust control for ancestry-related confounding but requires large sample sizes to detect rare variant associations¹⁸. To date, all rare variant association studies of ASD have been conducted using a family design, except for iPSYCH studies, which have leveraged population-based controls from national registers²⁸. Population stratification is a well-described confounder, especially for rare variants^{39,40}. However, effect estimates from family-based and unrelated case-control approaches were highly concordant for recurrent CNV, with no evidence of systematic bias. Similarly, the genome-wide analysis showed limited bias, further supported by the larger exome-based family analysis. Notably, the unrelated control analysis identified 6 times as many significant associations as the related control analysis, including 14 novel ASD loci. Overall, these results suggest that unrelated controls substantially increase power without introducing significant stratification bias, and do not support a significant role for indirect genetic effects in ASD liability, as such effects would be expected to inflate estimates with unrelated relative to related controls¹⁸. Together, these findings support the use of unrelated controls for rare variant discovery in ASD, complementing family-based designs for both CNVs and SNVs^{2,3}

ASD-associated loci overlap with other neurodevelopmental and psychiatric risk loci

We identified 2 recurrent CNVs, including duplications of 7q11 (*ZNF92*) and 9p24 (*DMRT1*), which were previously associated with schizophrenia but not ASD^{9,28}. In addition, among the new genome-wide candidate loci associated with autism, the 8q11 locus encompassing *RBICCI* has also been linked to schizophrenia when disrupted by CNVs as well as by SNVs^{33,41,42}. This locus-level pleiotropy has previously been observed for CNVs and rare gene-disrupting SNVs associated with schizophrenia and ASD^{34,43}. This stands in contrast to the modest overall genetic correlation between ASD and schizophrenia computed for common variants⁴⁴ ($r_g \sim 0.2$). While three of the 14 novel candidate loci have previously been associated with neurodevelopmental or psychiatric conditions^{23,31}, five additional loci have been identified in common-variant GWAS of other neurodevelopmental and psychiatric traits, such as ADHD⁴⁵, major depressive disorder⁴⁶, schizophrenia⁴⁷ and Tourette syndrome⁴⁸. Together, these findings suggest a convergence across both related disorders and distinct classes of genetic variants, with rare and common variants implicating shared genes.

Gene dosage effects on ASD liability and cognitive ability reflect partially distinct genetic architectures

The genetic overlap between autism and cognitive ability is an unresolved question. We observed a negative correlation between estimates of ASD liability and cognitive ability, suggesting that substantial liability for ASD is accompanied by substantial effects on cognition. However, we also observed several CNVs with a preferential effect on one trait over the other. The 9p24 duplication encompassing *DMRT1* (a transcription factor under genetic constraint) was associated with ASD and has previously been linked to schizophrenia⁹, but did not show any detectable effect on cognitive ability. This suggests that a duplication of a constrained gene may be associated with psychopathology with limited effects on cognitive ability. At the 1q21.1 distal locus, deletions and duplications showed preferential effects on cognitive ability and ASD, respectively. Interestingly, the opposite pattern was observed at the adjacent 1q21.1 TAR locus, where deletions showed preferential effects on ASD and duplications on cognitive ability.

At the genome-wide level, gene clusters preferentially associated with ASD were enriched for duplications, whereas those associated with both ASD and cognitive ability were enriched for deletions. This is consistent with previous observations suggesting that duplications associated with autism risk show milder risk for cognitive impairment compared to ASD-associated

deletions⁷. Interestingly, genes in ASD-preferential clusters showed lower intolerance to haploinsufficiency (LOEUF) than those associated with both traits or cognition alone. While the existence of autism-specific genes remains debated with conflicting results^{3,16,49}, our findings suggest that different classes of genetic variants with opposing effects on gene dosage (deletions versus duplications) may play a central role in shaping preferential phenotypic outcomes.

Pleiotropy across traits measured in ASD and general population cohorts.

Overall, deletions were more strongly correlated across developmental milestones and adult traits than duplications, suggesting greater pleiotropy, which is consistent with recent findings for CNVs in UKBB^{35,50}. In contrast, duplications appeared to be more specifically associated with psychopathology, with limited effects on cognitive ability as well as associated motor milestones such as the age of walking. Previous studies have sought to identify genes that may be preferentially associated with psychopathology rather than with cognition or global developmental delay^{2,3,16}. Those studies, however, were mainly focused on *de novo* loss-of-function variants. Our results suggest that classes of variants, rather than genes, may show preferential phenotypic associations.

Limitation (143 words)

Our unrelated case-control analysis design relies on external population-based cohorts for controls. We attempted to limit potential biases by excluding CNVs with heterogeneous frequencies across cohorts and batches, by replicating associations after restricting to European-ancestry participants, and by conducting multiple sensitivity analyses. Despite this very conservative approach, we cannot exclude spurious associations arising from residual population stratification, particularly for rare variants whose frequencies may vary across ancestries⁵¹.

Additionally, it is known that the UK Biobank is enriched for healthy volunteers⁵², which may inflate the estimates in the unrelated case-control analysis. Conversely, several CNVs previously associated with ASD (e.g., the 7q11.23 Williams-Beuren duplication) did not reach significance in either analysis. This may reflect changes in ASD diagnostic criteria and implementation over time, as individuals recruited in current autism cohorts tend to show less severe neurodevelopmental symptoms than those in earlier studies.

Conclusion (54 words)

This large CNV-GWAS (with a magnitude larger than in previous studies of ASD) demonstrates that incorporating controls from general population cohorts can substantially increase discovery without introducing major biases. The 14 novel candidate loci suggest that novel loci identified in future associations will predominantly involve ultra-rare variants or variants with smaller effect sizes.

Acknowledgments

This research was enabled by support provided by Calcul Quebec (<http://www.calculquebec.ca>) and the Digital Research Alliance of Canada (Alliance de recherche numérique du Canada). Sebastien Jacquemont is a recipient of a chair from the Jeanne et Jean Louis Levesque Foundation. This work is supported by a grant from the CIHR grant 159734 (Sebastien Jacquemont). Funding for the project was provided by the Wellcome Trust. This work was also supported by an NIH award U01 MH119690 granted to Laura Almasy, Sebastien Jacquemont and David Glahn and U01 MH119739.

We are grateful to all the families participating in this research, including SPARK, Simons Simplex Collection (SSC), MSSNG, Lothian Birth Cohort 1936 (LBC1936), CARTaGENE, Generation Scotland, IMAGEN, The Saguenay Youth Study (SYS), and UK Biobank Cohorts. We are grateful to all of the families in SPARK, the SPARK clinical sites and SPARK staff. We are also grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to phenotypic and genetic data for SPARK and SSC on SFARI Base. Approved researchers can obtain the SPARK and SSC population datasets described in this study by applying at <https://base.sfari.org/>. This research has been conducted using the UK Biobank Resource under Application Number 40980.

The Canadian Institutes of Health Research and the Heart and Stroke Foundation of Canada fund the Saguenay Youth Study (SYS). SYS was funded by the Canadian Institutes of Health Research and the Heart and Stroke Foundation of Canada. LBC1936 is supported by the Biotechnology and Biological Sciences Research Council and the Economic and Social Research Council (BB/W008793/1; which supports S.E.H.), Age UK (Disconnected Mind project), the Milton Damerel Trust, and the University of Edinburgh. S.R.C. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). Genotyping was funded by the BBSRC (BB/F019394/1). Generation Scotland received core support from the Chief Scientist Office of the Scottish Government

Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006] and is currently supported by the Wellcome Trust [216767/Z/19/Z]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award “STratifying Resilience and Depression Longitudinally” (STRADL) Reference 104036/Z/14/Z). Thank you to all the Generation Scotland participants. IMAGEN received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT-2007-037286); the Horizon 2020-funded ERC Advanced Grant “STRATIFY” (Brain network based stratification of reinforcement-related disorders) (695313); the Human Brain Project (HBP SGA 2, 785907, and HBP SGA 3, 945539); the Medical Research Council Grant “c-VEDA” (Consortium on Vulnerability to Externalizing Disorders and Addictions) (MR/N000390/1); the NIH (R01DA049238, A decentralized macro and micro gene-by-environment interaction analysis of substance use behavior and its brain biomarkers); the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London; the Bundesministerium für Bildung und Forschung (BMBF grants 01GS08152 and 01EV0711; Forschungsnetz AERIAL 01EE1406A and 01EE1406B; Forschungsnetz IMAC-Mind 01GL1745B); the Deutsche Forschungsgemeinschaft (DFG grants SM 80/7-2, SFB 940, TRR 265, and NE 1383/14-1); the Medical Research Foundation and Medical Research Council (grants MR/R00465X/1 and MR/S020306/1); and the NIH-funded ENIGMA (grants 5U54EB020403-05 and 1R56AG058854-01). Further support was provided by grants from the ANR (ANR-12-SAMA-0004 and AAPG2019 – GeBra), the ERA-Net NEURON (AF12-NEUR0008-01 – WM2NA and ANR-18-NEUR00002-01 – ADORe), the Fondation de France (00081242), the Fondation pour la Recherche Médicale (DPA20140629802), the Mission Interministérielle de Lutte-contre-les-Drogues-et-les-Conduites-Addictives (MILDECA), the Assistance-Publique-Hôpitaux-de-Paris and INSERM (interface grant), Paris-Sud University IDEX 2012, the Fondation de l’Avenir (grant AP-RM-17-013), the Fédération pour la Recherche sur le Cerveau, Science Foundation Ireland (16/ERC/3797), the NIH (Axon, Testosterone and Mental Health during Adolescence; RO1 MH085772-01A1), and NIH Consortium grant U54 EB020403, supported by a cross-NIH alliance that funds Big Data to Knowledge Centres of Excellence. The authors wish to acknowledge the resources of MSSNG (www.mss.ng), Autism Speaks, and The Center for Applied Genomics at The Hospital for Sick Children, Toronto, Canada. We also thank the participating families for their time and contributions to this database, as well as the generosity of the donors who supported this program. This research has been conducted using data from CARTaGENE (<https://cartagene.qc.ca/en>).

Ethics declarations

Ethics approval and consent to participate

This study was approved by the Research Ethics Committee of the Sainte Justine Research Center (approval number: 2021-2812) and was conducted in accordance with the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS2) and the US Code of Federal Regulations. Informed consent was obtained from all participants prior to enrollment, in accordance with each cohort's established participant consent framework.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no conflicts of interest.

Author Contributions Statement

Conceptualization: C.Poulain, G.H., S.J.;

Data curation: C.Poulain., Z.S., G.H., F.B., V.-R.B., A.L.;

Formal Analysis: C.Poulain., C.Proulx.;

Funding acquisition: S.J., L.A., D.C.G.;

Investigation: C.Poulain., C.Proulx.;

Methodology: C.Poulain., F.B., T.R., G.H.;

Project administration: G.H., S.J.;

Resources: C.Poulain., T.R., F.B., S.R.C., D.P., G.D., P.R., S.E.H., G.S., Z.P., T.P.;

Software: C.Poulain, F.B.;

Supervision: S.J., G.H., C.M.T.G.;

Validation: C.Poulain., S.J.;

Visualization: C.Poulain., S.J.;

Writing – original draft: C.Poulain., G.H., V.-R.B., S.J.;

Writing – review & editing: C.Poulain, T.R., V.-R.B., G.H., A.L., S.J., J.E., S.E.H., L.A., D.C.G., W.E., O.S., C.M.T.G.

METHOD (2949 words)

Table 1 describes the cohorts and individuals included in this study.

COHORTS

We analyzed data from 566,272 individuals across three autism cohorts and six general population cohorts (Table 1).

Autism spectrum disorder cohorts

Three family-based autism cohorts were included in this study: the Simons Simplex Collection (SSC)⁵³, Simons Foundation Powering Autism Research for Knowledge (SPARK)⁵⁴, and MSSNG⁵⁵. In total, after phenotypic and genotypic quality control, these cohorts comprise 37,701 cases with autism (probands = 29,696) and 52,034 family members, including siblings and parents of the affected proband(s).

Unselected population cohorts

We included six cohorts from the general population, used as controls: CARTaGENE (CaG)⁵⁶, IMAGEN⁵⁷; Generation Scotland (G-Scot)⁵⁸; Lothian Birth Cohort 1936 (LBC)^{59,60}; the Saguenay Youth Study (SYS)⁶¹ and the UK Biobank (UKBB) cohort⁶². These cohorts were previously pooled and studied in Huguet et al. 2024⁶³.

PHENOTYPES

Cognitive ability

Cognitive ability was assessed using nonverbal intelligence quotient (NVIQ) standard scores, as these subscales have been shown to more accurately reflect cognitive abilities in autistic individuals⁶⁴. Accordingly, NVIQ standard scores were derived in the SSC, MSSNG, SYS (children), and Imagen cohorts as detailed previously⁶⁵. Similarly, in SPARK, NVIQ scores were accessible from the participant's charts.

To quantify NVIQ in SPARK participants with parent-reported IQ, we assigned the median NVIQ score to participants within each bin who had NVIQ scores directly available in their charts. This mapping was derived from the distribution observed in 1,203 individuals from the full SPARK cohort. In LBC, we used the Moray House Test, a strong index of general cognitive

ability, converted to an IQ-type scale (corrected for age in days at completion) to match the IQ-based scale used in the other cohorts.

To obtain a robust measure that accounts for the variance in cognitive ability among general population participants from the UK Biobank, G-Scot, SYS (parents), and CaG cohorts, we computed a general intelligence factor (g-factor), using principal component analysis (PCA) on a range of cognitive tests, according to the methods detailed in Huguet et al¹⁵. In G-Scot, the g-factor was computed from Logical memory, Digit symbol, Verbal fluency, and Mill Hill Vocabulary tasks; in CaG, from Reasoning, Memory, and Reaction time tasks; and in SYS (parents), from 12 cognitive measures. For the UK Biobank, g-factors were computed from clinic-based assessments using five cognitive tasks, based on previous research⁶⁶, selecting the earliest available assessment measures for each participant: Fluid intelligence score (data field: 20016), Pairs matching (field: 399), Numeric memory (field: 4282), Prospective memory (field: 20018) and Reaction time (field: 20023). For online assessments, g-factor was derived from the following tasks at the first wave only: Fluid intelligence score (field: 20191), Trail making #2 (field: 20157), Symbol digit substitution (field: 20159), Pairs matching (field: 20132), Numeric memory (field: 20240). To maximize participant inclusion, we used the most robust available cognitive measure, following this priority: clinic g-factor, online g-factor, clinic Fluid Intelligence, and online Fluid Intelligence; if a g-factor could not be computed (due to incomplete cognitive data).

All cognitive ability measures were adjusted for age and sex and standardized within each cohort. This approach and the resulting composite cognitive scores have been described elsewhere⁶³.

Developmental phenotypes and autism-related traits

We extracted 16 traits from neurodevelopmental cohorts that have previously been associated with rare-variant burden scores^{7,67,68}. Developmental milestones were assessed based on parent report, across gross motor (ages at crawling, sitting, walking independently), fine motor (feeding oneself with a spoon), language (first words, first phrase), and toileting (bowel and bladder training). We included parent-reported professional diagnosis of intellectual disability, as well as 5 other parent-reported developmental disabilities: motor delay or developmental coordination disorder, language delay or language disorder, learning disability, social

(pragmatic) communication disorder, and speech articulation problems. We included scores from the Vineland Adaptive Behavior Scales (VABS) and the Developmental Coordination Disorder Questionnaire (DCDQ).

Traits and phenotypes from the general population

We extracted 45 binary and continuous complex traits from the UK Biobank (Table S9), selecting the earliest available assessment measures for each participant. These are well-studied heritable phenotypes that represent a broad spectrum of traits and medical conditions previously investigated using common variants and rare SNVs^{35,69,70}. We excluded outliers by considering values beyond ± 6 standard deviations (SD) and applied PHESANT⁷¹, an inverse rank normalization transformation (IRNT). This method involves ranking the continuous data and then transforming these ranks into quantiles of the standard normal distribution⁷². This approach has been described previously⁵⁰.

GENETICS

Genotype quality control

We used PLINK⁷³ software for processing and quality control. We kept only autosomal SNPs with minor allele frequency (MAF) $> 5\%$, probes providing genotypes that are not violating the Hardy-Weinberg equilibrium (threshold $< 1 \times 10^{-6}$) and probes with call rates $> 90\%$. We also checked for array duplicates, sex discordance, and family relationships for each participant. Finally, ancestries (principal component [PC] 1 to 10) were determined with KING⁷⁴, using the standard process defined on the website (<https://www.kingrelatedness.com>) and the 1000 Genomes as reference.

CNV calling

CNVs were called using PennCNV⁷⁵ and QuantiSNP⁷⁶, and filtered based on previously published methods^{15,63,77} available online (<https://github.com/JacquemontLab/MIND-GENESPARALLEL-CNV>). The following parameters were used for both algorithms: number of probes for CNV ≥ 3 , CNV size $\geq 1\text{Kb}$, likelihood scores ≥ 15 . CNVs detected by both algorithms were combined (CNVision⁵) to minimize the number of potential false discoveries. We defined all CNVs with less than 2 copies as deletions and all CNVs with more than 2 copies as duplications. After this merging step, an in-house algorithm based on CNV was applied to

concatenate adjacent CNVs of the same type into a single event, based on the following criteria: a) gap between CNVs ≤ 150 kb; b) size of the CNVs ≥ 1000 bp; and c) number of probes ≥ 3 .

Array QC filtering

We used the following quality-control criteria: call rate $\geq 95\%$; log R ratio-standard deviation < 0.35 ; B allele frequency-standard deviation < 0.08 ; |waviness factor| < 0.05 . Finally, we excluded participants carrying at least one CNV ≥ 10 Mb (N=431, a widely used QC threshold for CNVs^{15,35,63,77}) because such large CNVs are rarely observed in general population cohorts and are almost always present as mosaics and/or somatic CNVs that can not be analyzed together with germline CNVs.

CNV QC filtering

Only autosomal CNVs were considered for this study. CNVs with the following criteria were selected for analyses: likelihood score ≥ 30 (for at least one detection algorithm), size ≥ 50 kb, unambiguous type (deletions and duplications) and overlap with segmental duplications, HLA regions or centromeric regions $< 50\%$ (regions defined from UCSC hg19). In addition, we applied an in-house algorithm based on a machine learning method to detect additional artifact CNVs (DigCNV;<https://github.com/JacquemontLab/DigCNV>). This algorithm combines the consensus of three methods (Random forest, bagging of KNN and SVM) and uses 9 CNV features (Array criteria: log R ratio-standard deviation, B allele frequency-standard deviation, wave frequency; Localization CNV criteria: % of CNV overlap with centromeric regions and with segmental duplications; CNV criteria: density of SNPs (numbers of SNPs / size of CNV), likelihood score/number of SNPs, % algorithms overlapping, percentage of shared sequence found by both algorithms), type of CNV). The model was trained on 66% and tested on 33% of a manually curated reference set of 34,156 CNVs (31,746 true CNVs and 2,410 artifacts from 6 cohorts⁶³), validated using log R and BAF plots by two experts (a third in case of disagreement). DigCNV achieved an AUC of 0.95, a sensitivity of 0.95, and a specificity of 0.85.

Annotation of CNVs

CNVs were annotated using Gencode V19 annotation (the reference release for hg19 Human genome release) with ENSEMBL gene (<https://grch37.ensembl.org/index.html>). Each gene was annotated with “Loss-of-function observed/expected upper bound fraction” (LOEUF)⁷⁸.

We used the LOEUF score to measure intolerance to haploinsufficiency. As described by gnomAD, LOEUF is a continuous variable, ranging from 0 to 2, which provides insight across the full spectrum of Loss of Function constraint.

We defined recurrent CNVs (rCNV) as genomic regions previously associated with neurodevelopmental or neuropsychiatric conditions. The list was based on the ClinGen dosage sensitivity map, with additional loci with published evidence of clinical relevance (Table S1). A CNV is considered recurrent if it overlaps all genes in the geneset of a given rCNV. For each rCNV, the corresponding geneset was defined as protein-coding genes whose canonical transcripts overlapped by at least 50%. When a CNV matched multiple rCNVs, it was annotated using the rCNV with the largest gene set.

Mitigating batch effects from heterogeneous array platforms across cohorts

In large-scale studies integrating CNV data from multiple sources, genotyping platform and CNV detection sensitivity vary across cohorts, which can generate spurious associations. Here, we implemented a systematic approach to detect CNVs with heterogeneous frequencies across cohorts, batches, and array technologies.

First, to harmonize CNV detection sensitivity across array platforms, we only considered CNVs that were covered by at least 10 probes on every array platform. This strategy has been successfully applied in the past to investigate the effects of CNVs on cognitive ability^{14,63}. Second, to further control for batch-related differences, we identified and removed CNVs showing heterogeneous frequencies across cohorts, batches, and array technologies. We restricted this analysis to independent individuals in each cohort; for autism cohorts, we used parents rather than probands to maximize sample size. We considered 20 subgroups of individuals (Table S11), defined by cohort, genotyping technology, and batch. For each CNV, we computed frequencies within every subgroup and performed two chi-square tests for heterogeneity: one across the population-based subgroups and one across the autism-cohort subgroups. P-values were corrected for multiple testing using FDR correction within each set of tests. CNVs showing significant heterogeneity ($\text{FDR } q < 0.05$) in either test were excluded from further analyses unless they were mapped to a known recurrent CNV region.

This filtering step removed 55% of deletions and 68% of duplications. The vast majority of the removed CNVs were, on average, smaller, covered by fewer probes, and had lower likelihood

quality scores than retained CNVs. In addition, most of the removed CNVs were non-coding and would not have been included in downstream analyses. Of note, many of these CNVs represent recurrent events at the same genomic locus. When considering unique CNVs fully encompassing at least one gene, the removed fraction was lower: 22% of deletions and 29% of duplications (Figure S9).

CNV calling and filtering from Whole Exome Sequencing data

To increase power for the family-based analysis, we performed a sensitivity analysis using CNVs called from whole-exome sequencing data. We extracted copy number variations (CNVs) identified using xHMM⁷⁹ from the SPARK Release iWESv3 dataset⁵⁴. In accordance with the recommendations from the xHMM authors⁸⁰, we included all CNVs labelled as discovered and subsequently filtered out those with a quality score (SQ) < 90 and a Non-diploid quality (NDQ) < 99. To ensure consistency with CNVs identified through genotyping arrays, we applied analogous filtering criteria. We retained only autosomal CNVs with a maximum frequency < 1/1,000, either within our dataset or across all gnomAD ancestries, and ensured that less than 75% of their length overlapped problematic regions as defined by the UCSC Genome Browser. Given our focus on protein-coding genes affected by these CNVs, we implemented additional specific filters. We included only MANE genetic transcripts that were entirely encompassed within the CNVs and had less than 20% of their transcript length overlapping with problematic regions.

STATISTICAL ANALYSES

Analyses were performed using Python (3.13.2) with statsmodels (0.13.5) and R (4.1.2). Genomic visualizations were created using the Gviz⁸¹ R package.

Analytical design

We defined two complementary datasets: a family-based dataset comparing autistic individuals with unaffected relatives (N = 68,605; ASD = 25,888; Controls = 42,717), and an unrelated case-control dataset comparing autistic probands with individuals from unselected population cohorts, selecting one proband per family (N = 512,690; ASD = 29,696; Controls = 482,994). Statistical models within the family-based dataset accounted for familial relatedness.

Recurrent CNVs estimation with autism

We modelled autism risk for each recurrent CNV (defined based on the ClinGen dosage sensitivity map; Table S1), with at least 15 observations in our dataset, testing deletions and duplications separately. We performed the analyses in both the family-based and unrelated case-control datasets. Additionally, we tested a model for autism risk, adjusting for cognitive ability in the unrelated case-control dataset. We defined a significant Bonferroni threshold using the maximum number of recurrent CNVs tested across the 3 analyses ($n = 56$; Bonferroni $p < 8.9\text{e-}04$). Post-hoc power calculations, using the Bonferroni threshold defined above, indicated that 15 carriers provided 80% power to detect CNVs with large effect size ($\text{OR} \geq 7.0$; Figure S10).

CNV Genome-Wide Association Study

We fit a statistical model for each gene fully encompassed in a CNV with a minimum of 15 carriers among the individuals selected for the analysis. Post-hoc power calculations, using the FDR-corrected significance threshold from the ASD unrelated case-control analysis, indicated that this minimum carrier threshold provided 80% power to detect CNVs with large effect size ($\text{OR} \geq 5.2$; Figure S10). We used conditional logistic regression to compare autism cases with family-based (related) controls, and standard logistic regression with unselected controls. CNV-GWAS for continuous traits were analyzed using linear models, accounting for family effects when necessary. In each analysis, statistical models were applied separately for deletions and duplications. All models included age, sex, and 10 ancestry principal components (PCs) as covariates, except for autism and developmental milestones, which were adjusted for sex and PCs only. A false discovery rate (FDR) correction was applied to all models (deletions and duplications) for each phenotype separately.

Gene clusters

To cluster genes encompassed in multi-genic CNVs, we created gene clusters based on overlapping carrier counts, computed across all individuals in the study. For each pair of genes within a CNV, we computed the squared Pearson correlation coefficient (r^2). Let p_a and p_b denote the proportions of individuals carrying a CNV overlapping gene A and gene B, respectively. Similarly, p_{ab} represents the proportion of individuals carrying CNVs overlapping both genes A and B.

$$r^2 = \frac{(p_{AB} - p_A * p_B)^2}{p_A * (1 - p_A) * p_B * (1 - p_B)}$$

Pairs of genes with an $r^2 > 0.5$ were identified as connected pairs. Gene clusters were then constructed using a Depth-First Search (DFS) algorithm to group all correlated genes, ensuring that each gene belonged to a single cluster. In this framework, each cluster corresponds to a set of genes sharing at least 50% overlap across carriers, either through direct or transitive connectivity.

To define an r^2 threshold, we assumed that for a given recurrent CNV, a correctly defined cluster should include all and only the genes encompassed by that recurrent CNV. For each r^2 threshold, we computed, for every recurrent CNV, the sensitivity (the number of genes from the recurrent CNV captured within the corresponding cluster, divided by the total number of genes in the recurrent CNV) and the specificity (the number of genes from the recurrent CNV within the corresponding cluster, divided by the total number of genes in that cluster). Sensitivity and specificity were then averaged across all recurrent CNVs for each threshold. The r^2 threshold was selected as a compromise between average sensitivity and specificity, while also limiting the total number of clusters tested and increasing the proportion of multi-gene clusters (Figure S11).

Downsampling the unrelated analysis to the family-based sample size

To test whether the greater number of associations in the unrelated case-control analysis simply reflected its larger sample size, we downsampled the unrelated dataset to match the size and case-to-control ratio of the family-based analysis. We first counted the informative autism families contributing to the conditional logistic regression ($N = 23,070$), then derived the number of controls by applying the 1.65 case:control ratio of the main family-based analysis, yielding 23,070 ASD cases and 38,067 controls. We generated 500 bootstrap samples at this size and re-ran the ASD CNV-GWAS in each using only unrelated controls.

The number of significant genes and clusters (FDR $q < 0.05$) in the bootstrap samples was systematically higher than in the family-based analysis (empirical $p < 0.002$, Figure S3).

Defining novel robust associations with autism

Novel associations were defined among CNV-GWAS loci identified in the ASD unrelated case-control analysis, using the following criteria: (i) Significant after Bonferroni correction

based on the number of gene clusters tested ($p < 9.7\text{e-}06$); (ii) Replicated in the analysis restricted to individuals of European ancestry, using the Bonferroni threshold derived from the number of clusters tested in that analysis ($p < 1.1\text{e-}04$); (iii) Not overlapping with regions containing known rCNVs or high-confidence SFARI genes (categories S and I).

***De novo* analyses**

We computed *de novo* frequencies in two large independent datasets: Decipher and the diagnostic laboratory of CHUSJ, to evaluate the level of genetic constraint for our novel candidate regions in manually curated datasets. To obtain robust estimates of *de novo* events, we removed all CNVs larger than 3 MB. We set the expected probability of a *de novo* event to 1%, consistent with previously reported *de novo* CNV mutation rates in control populations^{82,83}, and assessed enrichment using two-sided binomial tests, with 95% confidence intervals computed using the Clopper-Pearson method.

Correlation of CNV-GWAS estimates across traits

We conducted multiple CNV-GWAS for 36 heritable phenotypes from the UK Biobank, as well as for 15 developmental phenotypes and autism-related traits. For each pair of traits, we computed the Pearson correlation between the median of gene-cluster estimates, separately for deletions and duplications. All correlations were FDR-corrected. To limit the influence of estimation noise from underpowered traits, we restricted this analysis to traits with at least 5 significant clusters ($\text{FDR } q < 0.05$).

Resource availability

Lead contact: For additional information, as well as requests regarding resources, please direct your inquiries to the lead contact:

Poulain C.: cecile.poulain@umontreal.ca

Jacquemont S.: sebastien.jacquemont@umontreal.ca

Data availability

The raw data from population samples and clinical cohorts are protected and are not available due to data privacy laws. Access can be obtained upon application. Approved researchers can obtain the SPARK and SSC population datasets described in this study by applying at <https://base.sfari.org/>.

The UK Biobank data are available upon request via: <https://www.ukbiobank.ac.uk/>
Data from the MSSNG project are available at <https://research.mss.ng/> under controlled access to protect participant privacy. The Lothian Birth Cohort 1936 data are available upon request via: <https://lothian-birth-cohorts.ed.ac.uk>
CARTaGENE data are available upon request via: <https://portail.cartagene.qc.ca/>.
Generation Scotland data are available upon request via: <https://genscot.ed.ac.uk/>
IMAGEN data are available upon request via: <https://www.imagen-project.org/>
Data from the Saguenay Youth Study are available upon request addressed to Zdenka Pausova and Tomas Paus.

Derived measures at the individual level used in this study are available upon request (S.J., sebastien.jacquemont@umontreal.ca). Summary statistics of gene-based CNV-GWAS and recurrent CNV estimates have been deposited on Zenodo: <https://doi.org/10.5281/zenodo.21938499>

Code availability

All original scripts have been deposited and are publicly available as of the date of publication on GitHub or Zenodo repositories: 1) Quality control and annotation of CNVs : <https://github.com/JacquemontLab/MIND-GENESPARALLELCNV> and <https://github.com/JacquemontLab/DigCNV>; 2) All scripts for the analyses and figures presented in this manuscript : <https://doi.org/10.5281/zenodo.21938316>

REFERENCES

1. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5TM, 5th Ed.* xliv, 947 (American Psychiatric Publishing, Inc., Arlington, VA, US, 2013). doi:10.1176/appi.books.9780890425596.
2. 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).
3. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568–584.e23 (2020).
4. Sanders, S. J. *et al.* A framework for the investigation of rare genetic disorders in neuropsychiatry. *Nat. Med.* **25**, 1477–1487 (2019).
5. Sanders, S. J. *et al.* Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism. *Neuron* **70**, 863–885 (2011).
6. Shanta, O. *et al.* A cross-disorder analysis of CNVs finds novel loci and dose-dependent relationships of genes to psychiatric traits. 2025.07.11.25331310 Preprint at <https://doi.org/10.1101/2025.07.11.25331310> (2025).
7. Douard, E. *et al.* Effect Sizes of Deletions and Duplications on Autism Risk Across the Genome. *Am. J. Psychiatry* **178**, 87–98 (2021).
8. Malhotra, D. & Sebat, J. CNVs: Harbinger of a Rare Variant Revolution in Psychiatric Genetics. *Cell* **148**, 1223–1241 (2012).
9. Marshall, C. R. *et al.* Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
10. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
11. Girirajan, S. *et al.* Relative Burden of Large CNVs on a Range of Neurodevelopmental Phenotypes. *PLoS Genet.* **7**, e1002334 (2011).
12. Moreno-De-Luca, D. *et al.* Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol. Psychiatry* **18**, 1090–1095 (2013).

13. Maenner, M. J. *et al.* Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveill. Summ.* **72**, 1–14 (2023).
14. Huguet, G. *et al.* Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. *JAMA Psychiatry* <https://doi.org/10.1001/jamapsychiatry.2018.0039> (2018) doi:10.1001/jamapsychiatry.2018.0039.
15. Huguet, G. *et al.* Genome-wide analysis of gene dosage in 24,092 individuals estimates that 10,000 genes modulate cognitive ability. *Mol. Psychiatry* <https://doi.org/10.1038/s41380-020-00985-z> (2021) doi:10.1038/s41380-020-00985-z.
16. Myers, S. M. *et al.* Insufficient Evidence for “Autism-Specific” Genes. *Am. J. Hum. Genet.* **106**, 587–595 (2020).
17. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
18. Davies, N. M. *et al.* The importance of family-based sampling for biobanks. *Nature* **634**, 795–803 (2024).
19. Howe, L. J. *et al.* Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nat. Genet.* **54**, 581–592 (2022).
20. Riggs, E. *et al.* Towards an evidence-based process for the clinical interpretation of copy number variation. *Clin. Genet.* **81**, 403–412 (2012).
21. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505**, 361–366 (2014).
22. Cooper, G. M. *et al.* A Copy Number Variation Morbidity Map of Developmental Delay. *Nat. Genet.* **43**, 838–846 (2011).
23. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
24. D’Angelo, D. *et al.* Defining the Effect of the 16p11.2 Duplication on Cognition, Behavior, and Medical Comorbidities. *JAMA Psychiatry* **73**, 20–30 (2016).

25. Jøneh, A. E. *et al.* Estimating the effect size of the 15Q11.2 BP1–BP2 deletion and its contribution to neurodevelopmental symptoms: recommendations for practice. *J. Med. Genet.* **56**, 701–710 (2019).
26. Jacquemont, S. *et al.* Genes To Mental Health (G2MH): A Framework to Map the Combined Effects of Rare and Common Variants on Dimensions of Cognition and Psychopathology. *Am. J. Psychiatry* **179**, 189–203 (2022).
27. Szatkiewicz, J. P. *et al.* Copy Number Variation in Schizophrenia in Sweden. *Mol. Psychiatry* **19**, 762–773 (2014).
28. Vaez, M. *et al.* Population-Based Risk of Psychiatric Disorders Associated With Recurrent Copy Number Variants. *JAMA Psychiatry* **81**, 957–966 (2024).
29. Foreman, J. *et al.* DECIPHER: Improving Genetic Diagnosis Through Dynamic Integration of Genomic and Clinical Data. *Annu. Rev. Genomics Hum. Genet.* **24**, 151–176 (2023).
30. Ruzzo, E. K. *et al.* Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850–866.e26 (2019).
31. Marshall, C. R. *et al.* Structural Variation of Chromosomes in Autism Spectrum Disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
32. Gonzalez-Mantilla, A. J., Moreno-De-Luca, A., Ledbetter, D. H. & Martin, C. L. A Cross-Disorder Method to Identify Novel Candidate Genes for Developmental Brain Disorders. *JAMA Psychiatry* **73**, 275–283 (2016).
33. Degenhardt, F. *et al.* Duplications in RB1CC1 are associated with schizophrenia; identification in large European sample sets. *Transl. Psychiatry* **3**, e326–e326 (2013).
34. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509–516 (2022).
35. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *Am. J. Hum. Genet.* **0**, (2022).
36. Crawford, K. *et al.* Medical consequences of pathogenic CNVs in adults: analysis of the UK Biobank. *J. Med. Genet.* **56**, 131–138 (2019).

37. Griswold, A. J. *et al.* Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum. Mol. Genet.* **21**, 3513–3523 (2012).
38. Pinto, D. *et al.* Functional Impact of Global Rare Copy Number Variation in Autism Spectrum Disorder. *Nature* **466**, 368–372 (2010).
39. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243–246 (2012).
40. Persyn, E., Redon, R., Bellanger, L. & Dina, C. The impact of a fine-scale population stratification on rare variant association test results. *PLoS ONE* **13**, e0207677 (2018).
41. Errichiello, E. *et al.* RB1CC1 duplication and aberrant overexpression in a patient with schizophrenia: further phenotype delineation and proposal of a pathogenetic mechanism. *Mol. Genet. Genomic Med.* **9**, e1561 (2020).
42. Singh, T. *et al.* Rare coding variants in 10 genes confer substantial risk for schizophrenia. *Nature* **604**, 509–516 (2022).
43. Rees, E. *et al.* Schizophrenia, autism spectrum disorders and developmental disorders share specific disruptive coding mutations. *Nat. Commun.* **12**, 5353 (2021).
44. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
45. Anney, R. J. L. *et al.* Conduct disorder and ADHD: Evaluation of conduct problems as a categorical and quantitative trait in the international multicentre ADHD genetics study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 1369–1378 (2008).
46. Nieratschker, V. *et al.* MORC1 exhibits cross-species differential methylation in association with early life stress as well as genome-wide association with MDD. *Transl. Psychiatry* **4**, e429–e429 (2014).
47. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
48. Yu, D. *et al.* Interrogating the genetic determinants of Tourette syndrome and other tic disorders through genome-wide association studies. *Am. J. Psychiatry* **176**, 217–227 (2019).

49. Wang, T. *et al.* Integrated gene analyses of de novo variants from 46,612 trios with autism and developmental disorders. *Proc. Natl. Acad. Sci.* **119**, e2203491119 (2022).
50. Kazem, S. *et al.* Determinants of pleiotropy and monotonic gene dosage responses across human traits. Preprint at <https://doi.org/10.21203/rs.3.rs-7547111/v1> (2025).
51. Schultz, L. M. *et al.* Copy-number variants differ in frequency across genetic ancestry groups. *Hum. Genet. Genomics Adv.* **5**, 100340 (2024).
52. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).
53. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. *Neuron* **68**, 192–195 (2010).
54. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).
55. Yuen, R. K. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
56. Awadalla, P. *et al.* Cohort profile of the CARTaGENE study: Quebec’s population-based biobank for public health and personalized genomics. *Int. J. Epidemiol.* **42**, 1285–1299 (2013).
57. Schumann, G. *et al.* The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. *Mol. Psychiatry* **15**, 1128–1139 (2010).
58. Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42**, 689–700 (2013).
59. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **41**, 1576–1584 (2012).
60. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1042r (2018).
61. Pausova, Z. *et al.* Cohort Profile: The Saguenay Youth Study (SYS). *Int. J. Epidemiol.* **46**, e19 (2017).

62. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
63. Huguet, G. *et al.* Effects of gene dosage on cognitive ability: A function-based association study across brain and non-brain processes. *Cell Genomics* **4**, (2024).
64. Dawson, M., Soulières, I., Ann Gernsbacher, M. & Mottron, L. The Level and Nature of Autistic Intelligence. *Psychol. Sci.* **18**, 657–662 (2007).
65. Douard, E. *et al.* Effects-sizes of deletions and duplications on autism risk across the genome. *Am. J. Psychiatry* **178**, 87–98 (2021).
66. Lyall, D. M. *et al.* Cognitive Test Scores in UK Biobank: Data Reduction in 480,416 Participants and Longitudinal Stability in 20,346 Participants. *PLOS ONE* **11**, e0154222 (2016).
67. Antaki, D. *et al.* A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. *Nat. Genet.* **54**, 1284–1292 (2022).
68. Warrier, V. *et al.* Genetic correlates of phenotypic heterogeneity in autism. *Nat. Genet.* **54**, 1293–1304 (2022).
69. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492–499 (2023).
70. Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 400,000 exomes. 2022.07.06.22277335 Preprint at <https://doi.org/10.1101/2022.07.06.22277335> (2022).
71. Millard, L. A., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
72. Jonsdottir, G. A. *et al.* Genetic propensities for verbal and spatial ability have opposite effects on body mass index and risk of schizophrenia. *Intelligence* **88**, 101565 (2021).
73. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–75 (2007).
74. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

75. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
76. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007).
77. Huguet, G. *et al.* Measuring and Estimating the Effect Sizes of Copy Number Variants on General Intelligence in Community-Based Samples. *JAMA Psychiatry* **75**, 447 (2018).
78. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
79. Fromer, M. *et al.* Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
80. Fromer, M. & Purcell, S. M. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **81**, 7.23.1–7.23.21 (2014).
81. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in *Statistical Genomics: Methods and Protocols* (eds Mathé, E. & Davis, S.) 335–351 (Springer, New York, NY, 2016). doi:10.1007/978-1-4939-3578-9_16.
82. Thapar, A. & Rutter, M. Genetic Advances in Autism. *J. Autism Dev. Disord.* **51**, 4321–4332 (2021).
83. Pinto, D. *et al.* Convergence of Genes and Cellular Pathways Dysregulated in Autism Spectrum Disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).

Figures

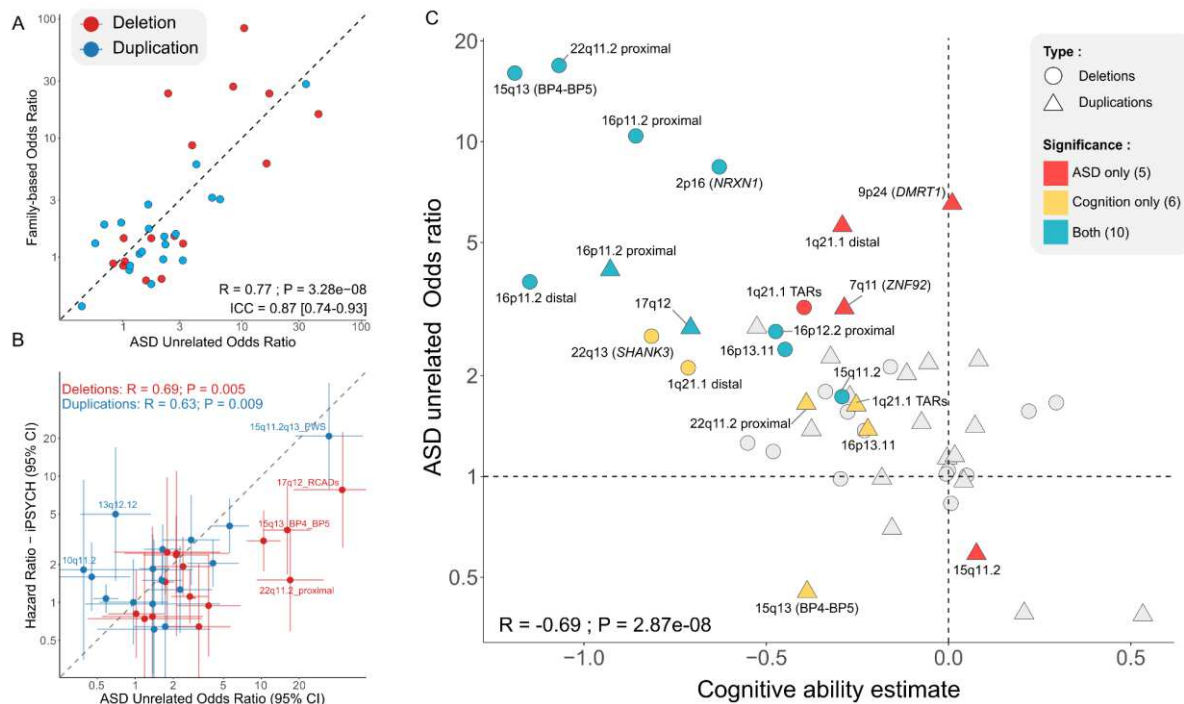


Figure 1. Effects of recurrent CNVs on autism risk and cognitive ability. (A) Comparison of ASD odds ratio estimates for recurrent CNVs between unrelated case-control and family-based analyses. All recurrent CNVs with > 15 carriers in both analyses are represented. (B) Comparison of ASD odds ratios and hazard ratios for recurrent CNVs between the unrelated case-control analysis and the population-based analysis from iPSYCH (Vaez et al., 2024). Because the event (CNV) is rare, HR and OR are expected to yield similar values (Symons & Moore, 2002). Pearson correlations are reported separately for deletions (red) and duplications (blue). (C) Comparison of recurrent CNV effect estimates on autism risk and cognitive ability using unrelated controls. Circles indicate deletions and triangles indicate duplications. Colours denote Bonferroni significance ($p < 8.93 \times 10^{-4}$): red, ASD only; yellow, cognitive ability only; cyan, both traits. Grey points did not reach significance for either trait. R = Pearson correlation coefficient; P = corresponding p-value; ICC: intraclass correlation coefficient

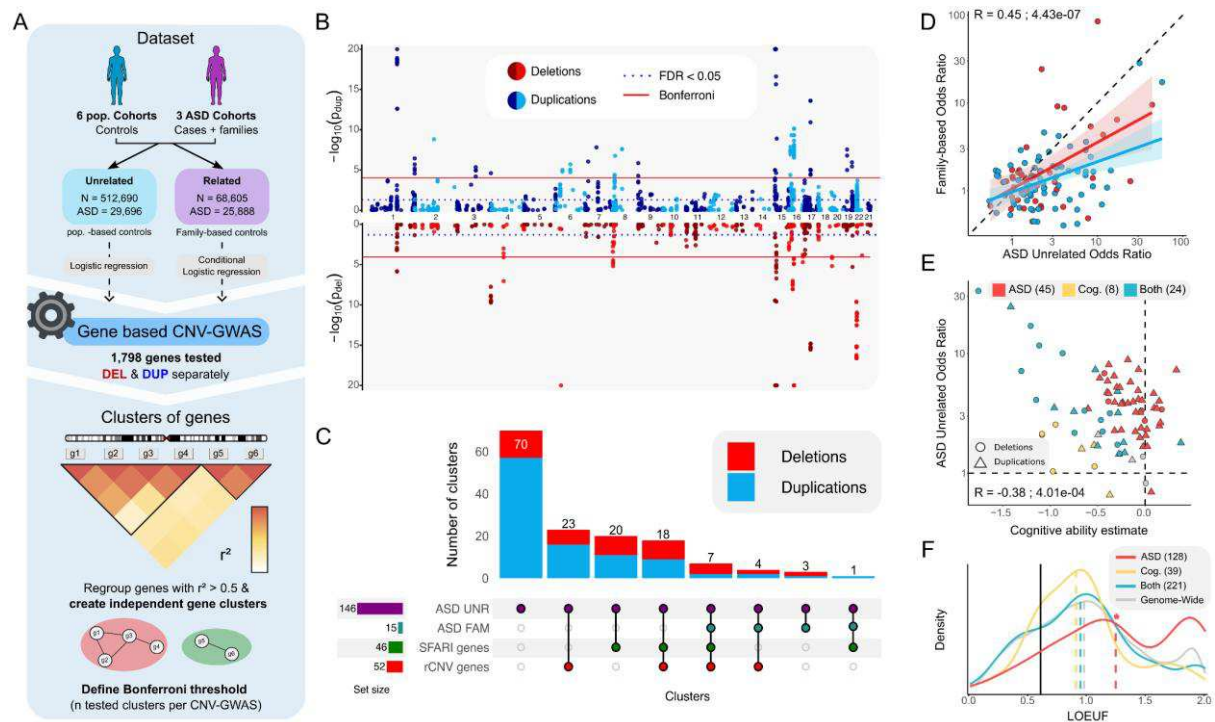


Figure 2. Genome-wide CNV associations with ASD and cognitive ability

(A) Workflow for the ASD CNV-GWAS. The unrelated and family-based analyses were performed separately. Gene-level associations were tested covering up to 1,798 genes (deletions and duplications separately), and genes with overlapping carriers ($r^2 > 0.5$) were grouped into independent clusters. **(B)** Miami plot of gene-level associations with ASD in the unrelated case-control analysis. The upper panel shows duplications, and the lower panel shows deletions ($-\log_{10}$ p-values). The dotted line indicates the FDR significance threshold ($q < 0.05$), and the solid red line indicates the Bonferroni-corrected threshold based on 530 independent gene clusters tested. **(C)** Upset plot showing the overlap of significant gene clusters (FDR $q < 0.05$) across ASD analyses. ASD UNR and ASD FAM represent clusters identified with unrelated case-control and family-based analysis, respectively. SFARI genes ($n = 46$) and rCNV genes ($n = 52$) denote clusters overlapping genes from the SFARI database or the recurrent CNV list, respectively. **(D)** Comparison of ASD odds ratio estimates for gene clusters between unrelated and family-based datasets. Red and blue dots represent deletions and duplications, respectively. **(E)** Comparison of gene cluster effect estimates between ASD risk and cognitive ability using unrelated controls. Circles indicate deletions and triangles indicate duplications. Colours denote FDR significance: red, ASD only ($n = 45$); yellow, cognitive ability only ($n = 8$); cyan, both traits ($n = 24$). Grey points did not reach significance for either trait. The Pearson correlation between the two sets of estimates is shown in the lower-left corner. **(F)** Distribution of LOEUF scores for genes within clusters associated either with ASD only (red, $n = 128$ genes), cognitive ability only (yellow, $n = 39$), or both traits (cyan, $n = 221$). The distribution of all genes with a LOEUF value is shown in grey. The asterisk (*) indicates a significant difference in median LOEUF relative to the genome-wide distribution (FDR-adjusted $p < 0.05$).

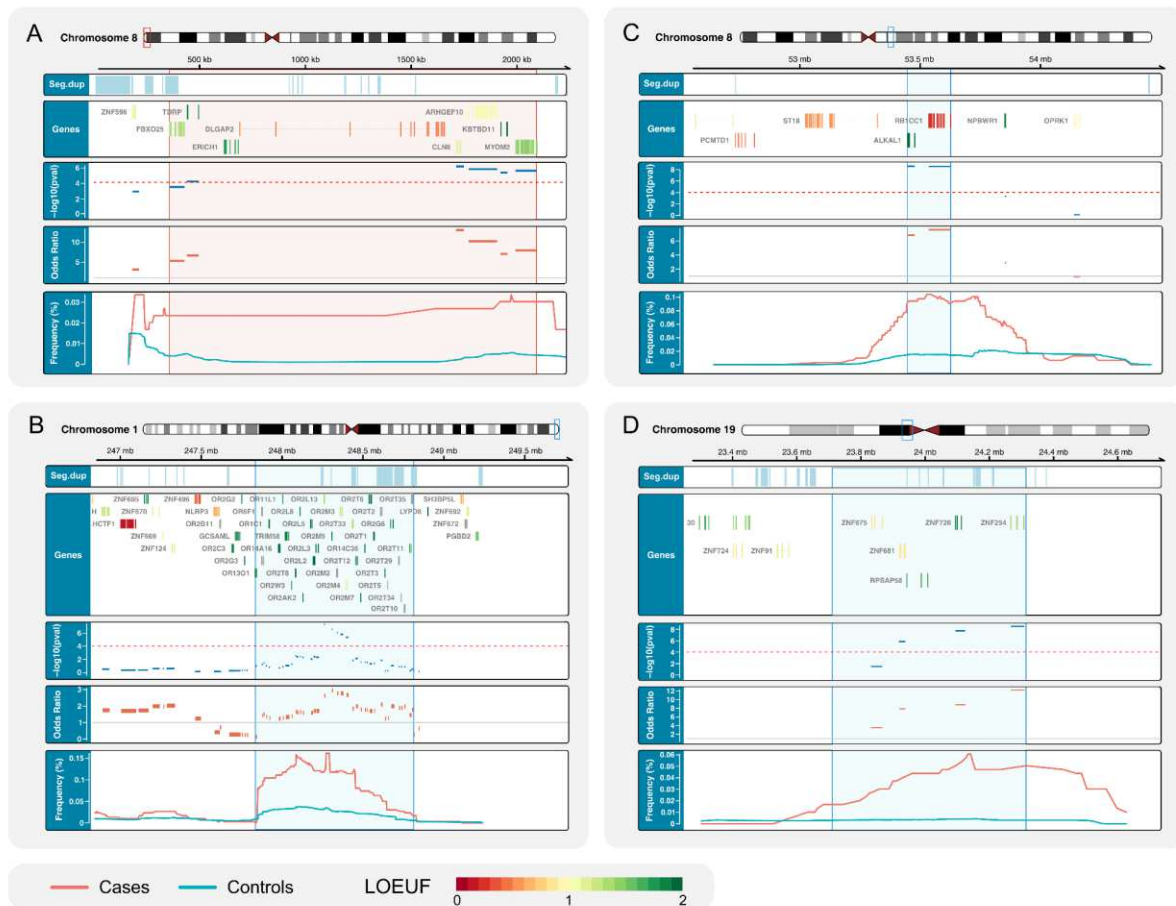


Figure 3. Novel ASD candidate loci with prior suggestive evidence

(A–D) Genomic visualization of four novel ASD-associated loci with prior suggestive evidence in the literature. For each locus, tracks display from top to bottom: segmental duplications, genes with their colour-coded constraint (LOEUF), gene-level association statistics ($-\log_{10}$ p-value), odds ratios, and CNV frequencies in cases (red) and controls (cyan). The dashed horizontal line indicates the Bonferroni significance threshold. Vertical lines denote the boundaries of the ASD-associated gene cluster. All association statistics are derived from the unrelated case-control analysis. (A) 8p23 deletion, containing SFARI category 2 genes *DLGAP2*, *CLN8*, and *ARHGEF10*. (B) 1q44 duplication, containing SFARI category 2 genes *OR2M4*, *OR1C1*, and *OR2T10*. (C) 8q11 duplication, containing *RB1CC1*, previously associated with schizophrenia. (D) 19p12 duplication, containing *ZNF675*, previously associated with developmental delay.

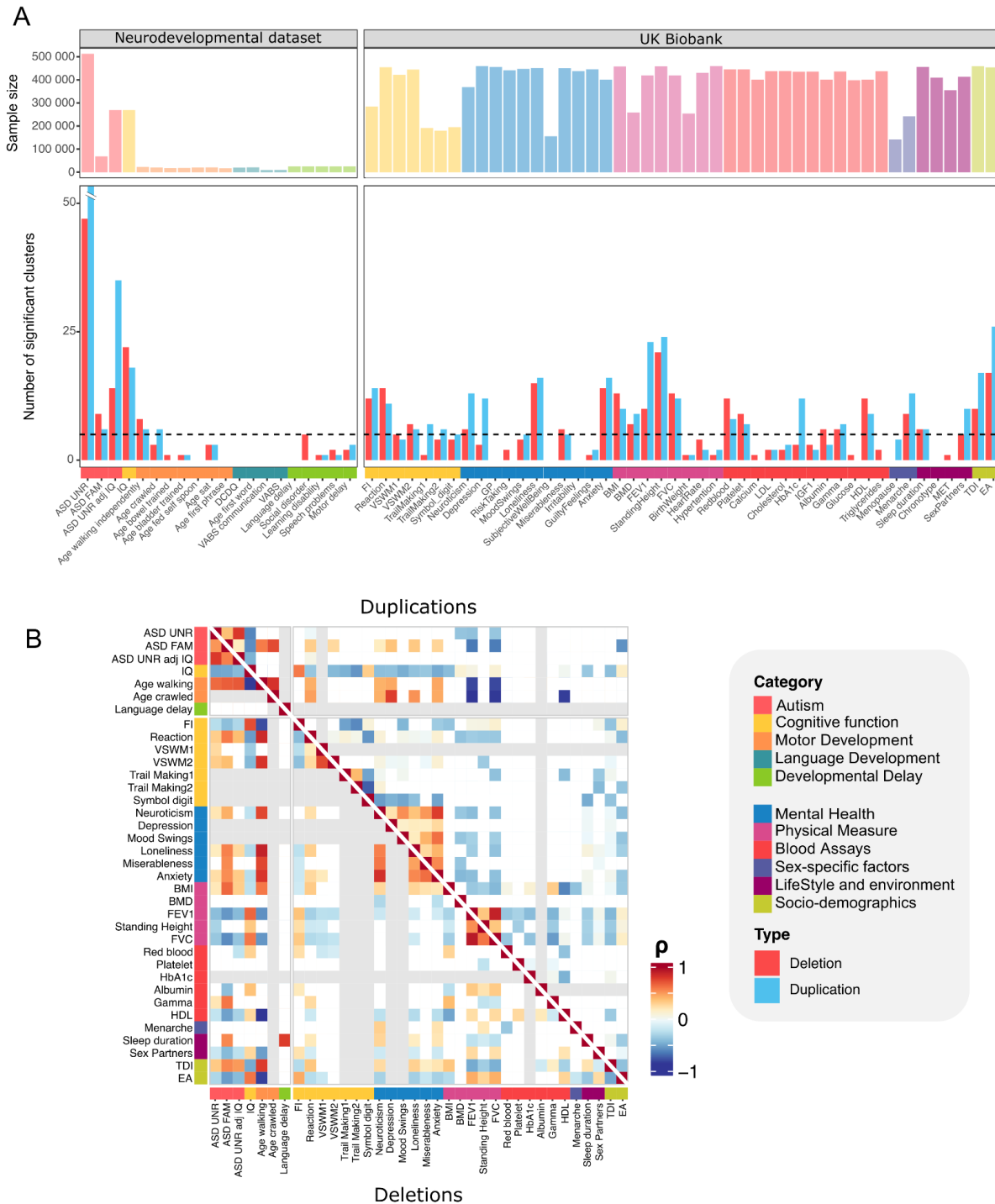


Figure 4. Gene dosage effects across neurodevelopmental and population-wide phenotypes

(A) Sample sizes (top) and number of significant gene clusters (FDR $q < 0.05$, bottom) for each CNV-GWAS conducted across phenotypes from the neurodevelopmental dataset and the UK Biobank. Bars are shown separately for deletions (red) and duplications (blue), and phenotypes are colour-coded by category. The dashed line indicates the threshold of five significant clusters, above which phenotypes were included in the correlation analysis in panel B. (B) Pairwise correlation matrix of gene cluster effect size estimates across traits with at least five

significant associations, shown separately for duplications (top) and deletions (bottom). Colours represent the direction of FDR-significant Pearson correlations. White cells indicate non-significant correlations; grey cells indicate trait pairs not tested due to fewer than five significant clusters. ASD UNR = ASD analysis with unrelated controls; ASD FAM = ASD analysis with family-based controls; ASD UNR adj IQ = ASD analysis with unrelated controls, adjusted for cognitive ability; FI = Fluid Intelligence score; VSWM1 = Visuospatial Working Memory (Number of incorrect matches); VSWM2 = Visuospatial Working Memory (Time to complete round); DCDQ = Developmental Coordination Disorder Questionnaire; VABS = Vineland Adaptive Behavior Scales standard score; TDI = Townsend Deprivation Index at recruitment; EA = Educational Attainment.

Symons, M. J., & Moore, D. T. (2002). Hazard rate ratio and prospective epidemiological studies. *Journal of Clinical Epidemiology*, 55(9), 893–899. [https://doi.org/10.1016/S0895-4356\(02\)00443-2](https://doi.org/10.1016/S0895-4356(02)00443-2)

Vaez, M., Montalbano, S., Calle Sánchez, X., Georgii Hellberg, K.-L., Dehkordi, S. R., Krebs, M. D., Meijsen, J., Shorter, J., Bybjerg-Grauholm, J., Mortensen, P. B., Børglum, A. D., Hougaard, D. M., Nordentoft, M., Geschwind, D. H., Buil, A., Schork, A. J., Helenius, D., Raznahan, A., Thompson, W. K., ... iPSYCH Investigators. (2024). Population-Based Risk of Psychiatric Disorders Associated With Recurrent Copy Number Variants. *JAMA Psychiatry*, 81(10), 957–966. <https://doi.org/10.1001/jamapsychiatry.2024.1453>

Tables

Cohort	Group	Entire Dataset				Cognitive Subset			
		N proband	N total	Male %	Mean Age	N proband	N total	Male %	Mean Age
Unselected Cohorts									
Cag	Controls	-	5,763	41%	54.85	-	2,600	47%	53.94
G-Scot	Controls	-	14,159	41%	46.85	-	13,744	41 %	46.72
Imagen	Controls	-	1,787	48%	14.45	-	1,742	48%	14.45
LBC	Controls	-	554	51%	70.00	-	504	51%	70.00
SYS	Controls	-	1,869	48%	28.24	-	1,569	47%	28.24
UKBB	Controls	-	459,870	45%	58.45	-	238,348	45%	58.45
Autism Cohorts									
MSSNG	Cases	2,158	2,541	78%	9.43	898	1,046	79%	9.44
MSSNG	Controls	-	2,686	49%	49.87	-	96	50%	40.24
SPARK	Cases	24,957	32,579	76%	12.00	8,228	10,545	78%	11.59
SPARK	Controls	-	34,315	38%	36.50	-		-	-
SSC	Cases	2,581	2,581	86%	09.03	2,565	2,565	86%	09.03
SSC	Controls	-	7,568	49%	32.02	-	6	83%	47.56
Total (N)			566,272				272,265		

Table 1. Descriptive characteristics of study cohorts

Summary of the 566,272 individuals included after quality control, drawn from six general population cohorts: CARTaGENE (CaG), Generation Scotland (G-Scot), the Saguenay Youth Study (SYS), the Lothian Birth Cohort 1936 (LBC), and the UK Biobank (UKBB); and three autism-specific cohorts: the Simons Simplex Collection (SSC), Simons Foundation Powering Autism Research for Knowledge (SPARK), and MSSNG. Probands refer to independent cases, while total cases additionally include relatives with an autism diagnosis. A subset of cognitive abilities is also described for analyses that require cognitive measures. Cognitive measures were extracted from Huguet et al. (2024).

CNV	Extra-familial dataset			Intra-familial dataset			Lit.OR
	carriers (ASD)	OR	P-value	carriers (ASD)	OR	P-value	
1q21.1 (TARs) Del.	105(15)	3.20[1.79-5.72]	8.7e-05*	19(10)	1.30[0.33-5.11]	7.1e-01	-
1q21.1 distal Del.	126(13)	2.11[1.16-3.83]	1.4e-02	26(10)	0.66[0.19-2.24]	5.0e-01	3,2
1q21.1 distal Dup.	199(47)	5.60[3.93-7.98]	1.5e-21*†	68(41)	3.16[1.58-6.31]	1.1e-03	5,3
2p16 (NRXN1) Del.	203(68)	8.41[6.11-11.58]	6.2e-39*†	124(76)	27.00[8.03-90.77]	9.9e-08*	7,9
3q29 (DLG1) Del.	18(11)	26.90[9.94-72.75]	8.9e-11*	11(10)	-	-	19
7q11 (ZNF92) Dup.	229(38)	3.18[2.19-4.62]	1.2e-09*†	84(30)	0.94[0.51-1.72]	8.4e-01	-
7q11.23 WBs Dup.	20(4)	3.75[1.08-13.02]	3.8e-02	7(5)	-	-	32
9p24 (DMRT1) Dup.	48(13)	6.53[3.30-12.92]	7.1e-08*†	22(11)	3.06[0.81-11.54]	9.9e-02	-
15q11.2 Del.	1951(177)	1.73[1.47-2.03]	2.0e-11*	390(165)	1.44[1.06-1.96]	2.1e-02	1,3
15q11.2 Dup.	2209(83)	0.59[0.47-0.74]	5.6e-06*	190(81)	1.30[0.86-1.97]	2.1e-01	-
15q11.2q13 PWS Dup.	52(34)	34.31[18.49-63.69]	3.9e-29*	32(28)	28.35[6.09-131.91]	2.0e-05*	-
15q13, BP3-BP4 Del.	28(9)	6.66[2.85-15.55]	1.2e-05*	13(8)	-	-	-
15q13, BP4-BP5 Del.	99(47)	16.04[10.38-24.77]	6.5e-36*†	47(32)	6.11[1.85-20.10]	2.9e-03	15
16p13.11 Del.	165(18)	2.40[1.44-3.99]	7.6e-04*	23(15)	23.67[4.57-122.43]	1.6e-04*	2,5
16p12.2 proximal Del.	317(43)	2.71[1.93-3.81]	8.2e-09*	73(37)	1.51[0.74-3.08]	2.6e-01	-
16p11.2 distal Del.	79(16)	3.81[2.12-6.85]	7.5e-06*	22(16)	8.68[2.20-34.22]	2.0e-03	3,6
16p11.2 proximal Del.	201(77)	10.40[7.60-14.24]	1.5e-48*†	61(51)	83.66[11.01-635.44]	1.9e-05*	14,3
16p11.2 proximal Dup.	120(24)	4.13[2.57-6.66]	5.5e-09*†	25(18)	6.01[1.85-19.47]	2.8e-03	10,5
17p11.2 Dup.	2(0)	-	-	-	-	-	32
17q12 (HNF1B) Del.	27(19)	43.66[17.98-106.01]	7.2e-17*	19(14)	15.89[3.10-81.53]	9.2e-04	3,5
17q12 (HNF1B) Dup.	123(18)	2.78[1.65-4.68]	1.2e-04*	40(20)	1.56[0.56-4.36]	3.9e-01	-
22q11.2 proximal Del.	49(24)	16.90[9.19-31.09]	9.9e-20*†	26(23)	23.62[2.98-187.35]	2.8e-03	32,3
22q11.2 proximal Dup.	442(40)	1.65[1.18-2.31]	3.8e-03	70(35)	1.73[0.83-3.60]	1.4e-01	2

Table 2. Associations between autism and recurrent CNVs
ASD association estimates for recurrent genomic loci in both extra-familial and intra-familial datasets. An additional extra-familial analysis adjusting for cognitive ability was conducted; CNVs that remained significant after adjustment are marked with (†). Bonferroni-significant associations (56 tests; $p < 8.93 \times 10^{-4}$) are shown, along with CNVs previously reported to be

associated with ASD (Jacquemont et al., 2022). The complete set of estimates is provided in Table S2.

Gene Cluster					Extra-familial dataset			De novo rate	
Type	Cytoband	N	Highlighted genes	$\Sigma 1/\text{LOEUF}$	carriers	OR	P-value	ratio	<i>de novo</i>
Del.	4q13	17	<i>OPRPN</i>	10.67	25	14.24	7.5e-09	1/10	0.10 [0.002 - 0.45]
Del.	8p23	6	<i>TDRP, CLN8, ARHGEF10, KBTBD11, MYOM2, (DLGAP2)</i>	7.28	18 - 30	8.24 [6.93 - 13.36]	2.3e-06	16/35	0.46 [0.29 - 0.63]*
Del.	16q23	11	<i>C16orf46, GCSH</i>	14.85	17 - 18	10.54 [10.11 - 10.96]	1.7e-05	2/11	0.18 [0.02 - 0.52]*
Del.	19q13	8	<i>NLRP2</i>	6.30	40	6.83	5.0e-07†	3/6	0.5 [0.12 - 0.88]*
Dup.	1p22	5	<i>BRDT, EPHX4</i>	12.28	39 - 40	5.77 [5.65 - 5.89]	3.0e-05	1/8	0.13 [0.003 - 0.53]
Dup.	1q44	38	<i>OR2L13, OR2M5, (OR2M4, OR1C1, OR2T10)</i>	31.02	160 - 207	2.92 [2.75 - 3.08]	1.1e-07	0/11	0.00 [0.00 - 0.28]
Dup.	3q13	3	<i>GUCA1C</i>	2.41	55	5.16	4.3e-07†	0/5	0.00 [0.00 - 0.52]
Dup.	6p21	4	<i>KCNK17, KCNK16</i>	7.07	15 - 17	14.84 [13.42 - 16.25]	1.4e-06	0/0	-
Dup.	6q14	5	<i>RIPPLY2, CYB5R4, MRAP2, CEP162</i>	22.94	20 - 25	11.28 [10.17 - 14.93]	9.4e-07	0/2	0.00 [0.00 - 0.84]
Dup.	7p21	2	<i>AGR3</i>	2.17	48	6.41	8.3e-08†	1/6	0.17 [0.004 - 0.64]
Dup.	8p22	2	<i>C8orf48</i>	2.11	94	4.64	1.5e-08†	1/10	0.1 [0.002 - 0.45]
Dup.	8q11	2	<i>ALKAL1, RB1CC1</i>	7.85	49 - 57	7.32 [6.92 - 7.72]	2.0e-09†	1/16	0.06 [0.001 - 0.30]
Dup.	19p12	4	<i>ZNF726, ZNF254</i>	7.19	30 - 35	10.64 [8.98 - 12.30]	8.1e-09	0/4	0.00 [0.00 - 0.60]
Dup.	19q13	9	<i>FPR3</i>	7.65	251	2.78	1.0e-07†	0/5	0.00 [0.00 - 0.52]

Table 3. Novel genome-wide CNV associations with ASD

Novel ASD-associated loci identified using a Bonferroni threshold ($p < 9.4\text{e-}05$), excluding regions overlapping recurrent CNVs or high-confidence SFARI genes (categories S and I), and replicated in the European ancestry analysis. None of these regions reached significance in the intra-familial dataset. N indicates the number of genes within the region. Highlighted genes represent the most significant genes within each

cluster; bold indicates SFARI genes within the region, and genes in parentheses are SFARI genes not driving the association signal. $\sum 1/\text{LOEUF}$ represents the sum of inverse LOEUF scores (V4) across genes within the cluster, reflecting the overall intolerance to loss-of-function variation. Regions remaining significant after adjustment for cognitive ability are marked with (†). De novo CNV frequencies were estimated in the DECIPHER and Sainte-Justine diagnostic databases using a binomial test with an expected de novo probability of 1%, and 95% confidence intervals were computed using the Clopper-Pearson method.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigures.docx](#)
- [Supplementarytables.xlsx](#)