nature portfolio

https://doi.org/10.1038/s41588-022-01104-0

Supplementary information

Rare coding variation provides insight into the genetic architecture and phenotypic context of autism

In the format provided by the authors and unedited

Supplementary Note

Autism Sequencing Consortium (ASC)

Branko Aleksic 50, Joon-Yong Anney 83, Richard Anney 84, Mykyta Artomov 1,2,4,69, Mafalda Barbosa 15,18, Elisa Benetti 34,35, Catalina Betancur 19, Monica Biscaldi-Schafer 59, Somer Bishop 9, Anders D. Børglum 51,52,53,54, Harrison Brand 1,2,3,7, Michael S. Breen 13,14,15, Alfredo Brusco 20,21, Joseph D. Buxbaum 13,15,46, Gabriele Campos 32, Simona Cardaropoli 26, Diana Carli 26, Angel Carracedo 57,70, Marcus C.Y. Chan 22, Andreas G. Chiocchetti 59, Brian H.Y. Chung 22, A. Ercument A. Cicek 49,85, Rachel Cohen 13,14, Brett Collins 13,14,15, Ryan L. Collins 1,2,3,8, Edwin H. Cook 23, Hilary Coon 55,56, Claudia I.S. Costa 32, Michael L. Cuccaro 24, David J. Cutler 44, Bernardo Dalla Bernardina 86, Mark J. Daly 1,2,4,5,47,48, Silvia De Rubeis 13,14,15,45, Bernie Devlin 11, Caroline Dias 71,87, Ryan N. Doan 71, Enrico Domenici 25, Shan Dong 9, Chiara Fallerini 34,35, Montserrat Fernández-Prieto 57,58, Giovanni Battista Ferrero 26, Christine M. Freitag 59, Menachem Fromer 2, Jack M. Fu 1,2,3, Louise Gallagher 88, J. Jay Gargus 27, Sherif Gerges 1,2,4,69, Daniel Geschwind 77,89, Michael Gill 88, Michael Gilson 9, Elisa Giorgio 20, Ana Cristina Girardi 32, Javier González-Peñas 65, Jakob Grove 51,52,53,54, Elizabeth E. Guerrero 29, Stephen Guter 23, Danielle Halpern 13,14, Emily Hansen-Kiss 41, Xin He 90, Gail E. Herman 28, Irva Hertz-Picciotto 29, David M. Hougaard 51,60, Christina M. Hultman 17, Iuliana Ionita-Laza 91, Suma Jacob 23, Jesslyn Jamison 13,14, Magdalena Janecka 13,14,92, Astanand Jugessur 81, Miia Kaartinen 66, Lambertus Klei 11, Gun Peggy Knudsen 81, Alexander Kolevzon 13,14,61, Itaru Kushima 50,72, So Lun Lee 22, Terho Lehtimäki 73, Tess Levy 13,14, Yue Li 6, Lindsay Liang 9, Elaine T. Lim 69, Carla Lintas 42, W. Ian Lipkin 91, Alicia Ljungdahl 9, Caterina Lo Rizzo 35,36, Fátima Lopes 30, Yunin Ludena 29, Patricia Maciel 30, Per Magnus 81, Behrang Mahjani 13,14,17, Nell Maltman 23, Marianna Manara 35,36, Dara S. Manoach 31, Gal Meiri 74,75, Idan Menashe 62,63, Judith Miller 76,77, Nancy Minshew 11, Eric M. Morrow 93, Matthew Mosconi 78, Pierandrea Muglia 94, Benjamin M. Neale 2,4,5, Rachel Nguyen 27, Utku Norman 85, Norio Ozaki 50,79, Aarno Palotie 2,5,48,64, Mara Parellada 65, Maria Rita Passos-Bueno 32, Lisa Pavinato 20, Minshi Peng 6, Margaret Pericak-Vance 24, Antonio M. Persico 33, Isaac N. Pessah 29,43, Kaija Puura 66, Abraham Reichenberg 13,14,15,80, Alessandra Renieri 34,35,36, Elise B. Robinson 2,4,5, Kathryn Roeder 6,49, Kaitlin E. Samocha 1,2, Stephan J. Sanders 9, Sven Sandin 13,14,17, Susan L. Santangelo 39,95, F. Kyle Satterstrom 2,4,5, Gerry Schellenberg 96, Stephen W. Scherer 67,68, Sabine Schlitt 59, Rebecca J. Schmidt 29, Lauren Schmitt 23, Katja Schneider-Momm 59, Tarjinder Singh 2,4,5,69,97, Paige M. Siper 13,14,15, Laura Sloofman 13,14,15, Moyra Smith 27, Gabriela Soares 98, Matthew W. State 9, Christine R. Stevens 2,4,5, Camilla Stoltenberg 81, Pål Suren 81, Ezra Susser 99,100, James S. Sutcliffe 37,38, John A. Sweeney 82, Peter Szatmari 101, Michael E. Talkowski 1,2,3,4,8, Flora Tassone 29,39, Karoline Teufel 59, Audrey Thurm 102, Elisabetta Trabetti 40, Slavica Trajkova 20, Maria del Pilar Trelles 13,14, Christopher A. Walsh 103,104,105, Brie Wamsley 10, Jaqueline Y.T. Wang 32, Yuting Wei 106, Lauren A. Weiss 9, Donna Werling 107, Emilie M. Wigdor 2,4,5, Jeremy A. Willsey 9,108,109, Mullin H.C. Yu 22, Timothy W. Yu 110, Ryan Yuen 67, Michael E. Zwick 44

iPSYCH-BROAD Consortium

Esben Agerbo 51,111,112, Thomas Damm Als 51,113,114, Vivek Appadurai 51, Marie Bækved-Hansen 51,115, Rich Belliveau 4, Carsten Bøcker Pedersen 51,111,112, Anders D. Børglum 51,52,53,54, Alfonso Buil 51,116, Jonas Bybjerg-Grauholm 51,60, Caitlin E. Carey 2,4,5, Felecia Cerrato 4, Kimberly Chambert 4, Claire Churchhouse 2,4,5, Søren Dalsgaard 51,113,114, Mark J. Daly 1,2,4,5,47,48, Ditte Demontis 51,113, Ashley Dumont 4, Mads Engel Hauberg 51,113,114, Marianne Giørtz Pedersen 51,111,112, Jacqueline Goldstein 2,4,5, Jakob Grove 51,52,53,54, Christine S. Hansen 51,60,116, Mads V. Hollegaard 60, David M. Hougaard 51,60, Daniel P. Howrigan 4,5, Hailiang Huang 4,5, Julian Maller 4,117, Alicia R. Martin 2,4,5, Joanna Martin 4,17,118, Manuel Mattheisen 51,113,119,120, Jennifer Moran 4, Ole Mors 51,121, Preben Bo Mortensen 51,52,111,112, Benjamin M. Neale 2,4,5, Merete Nordentoft 51,122, Jonatan Pallesen 51,113,114, Duncan S. Palmer 4,5, Timothy Poterba 2,4,5, Jesper Buchhave Poulsen 51,60, Stephan Ripke 2,5,123, Elise B. Robinson 2,4,5, F. Kyle Satterstrom 2,4,5, Andrew J. Schork 116, Christine R. Stevens 2,4,5, Wesley K. Thompson 116,124, Patrick Turley 4,5, Raymond K. Walters 4,5, Thomas Werge 51,116,125, Emilie M. Wigdor 2,4,5

Broad Institute Center for Common Disease Genomics (Broad-CCDG)

Felecia Cerrato 4, Claire Churchhouse 2,4,5, Caroline Cusick 4, Mark J. Daly 1,2,4,5,47,48, Anne Feng 126,127, Stacey B. Gabriel 16, Namrata Gupta 2, Sekar Kathiresan 1,104,128, Amit Khera 1, Eric Lander 104, Robert Maier 2,4,5, Benjamin M. Neale 2,4,5, Candace Patterson 2, Christine R. Stevens 2,4,5, Michael E. Talkowski 1,2,3,4,8

Affiliations

1. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 3. Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; 4. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 5. Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 6. Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA; 7. Pediatric Surgical Research Laboratories, Department of Surgery, Massachusetts General Hospital, Boston, MA 02114, USA; 8. Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA; 9. Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA 94143, USA; 10. Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095, USA; 11. Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA; 12. Data Sciences Platform, The Broad Institute of MIT and Harvard,Cambridge, MA 02142, USA; 13. Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 14. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 15. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 16. Genomics Platform, The Broad Institute of MIT

and Harvard, Cambridge, MA 02142, USA; 17. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden; 18. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 19. Sorbonne Université, INSERM, CNRS, Neuroscience Paris Seine, Institut de Biologie Paris Seine, 75005 Paris, France; 20. Department of Medical Sciences, University of Torino, Turin 10126, Italy; 21. Medical Genetics Unit, "Città della Salute e della Scienza" University Hospital, Turin 10126, Italy; 22. Department of Pediatrics & Adolescent Medicine, Duchess of Kent Children's Hospital, The University of Hong Kong, Hong Kong Special Administrative Region 999077, China; 23. Institute for Juvenile Research, Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60608, USA; 24. The John P Hussman Institute for Human Genomics, The University of Miami Miller School of Medicine, Miami, FL 33136, USA; 25. Department of Cellular, Computational and Integrative Biology, University of Trento, 38122 Trento, Italy; 26. Department of Public Health and Pediatrics, University of Torino, Turin 10126, Italy; 27. Center for Autism Research and Translation, University of California Irvine, Irvine, CA 92697, USA; 28. The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA; 29. MIND (Medical Investigation of Neurodevelopmental Disorders) Institute, University of California Davis, Davis, CA 95616, USA; 30. Life and Health Sciences Research Institute, School of Medicine, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal; 31. Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; 32. Centro de Pesquisas sobre o Genoma Humano e Células tronco, Instituto de Biociências, Universidade de São Paulo, São Paulo, 05508090, Brazil; 33. Child & Adolescent Neuropsychiatry Program, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, I-41125 Modena, Italy; 34. Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Ital; 35. Medical Genetics, University of Siena, 53100 Siena, Italy; 36. Genetica Medica, Azienda Ospedaliera Universitaria Senese, 53100 Siena, Italy; 37. Department of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; 38. Vanderbilt Genetics Institute, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; 39. Department of Biochemistry and Molecular Medicine, University of California Davis, School of Medicine, Sacramento, CA 95817, USA; 40. Department of Neurosciences, Biomedicine and Movement Sciences, Section of Biology and Genetics, University of Verona, 37134 Verona, Italy; 41. Department of Diagnostic & Biomedical Sciences, University of Texas Health Science Center at Houston, School of Dentistry, Houston, TX 77054, USA; 42. Service for Neurodevelopmental Disorders, University Campus Bio-medico of Rome, 00128 Rome, Italy; 43. Department of Molecular Biosciences, University of California Davis, School of Veterinary Medicine, Davis, CA 95616, USA; 44. Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322 USA; 45. Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 46. Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 47. Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; 48. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00290 Helsinki, Finland.; 49. Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA; 50. Department of Psychiatry, Graduate School of Medicine, Nagoya University, Nagoya 466-8550, Japan; 51. The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, 8000 Aarhus, Denmark; 52. Center for Genomics and Personalized Medicine, 8000 Aarhus, Denmark; 53. Department of Biomedicine - Human Genetics, Aarhus University, 8000 Aarhus, Denmark; 54. Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark; 55. Department of Internal Medicine, University of Utah, Salt Lake City, UT 84132, USA; 56. Department of Psychiatry, Huntsman Mental Health Institute, University of Utah, Salt Lake City, UT 84108, USA; 57. Grupo de Medicina Xenómica, Centro de Investigación en Red de Enfermedades Raras (CIBERER), CIMUS, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain; 58. Neurogenetics group, Instituto de Investigación Sanitaria de Santiago (IDIS-SERGAS), 15706 Santiago de Compostela, Spain; 59. Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe University Frankfurt, 60528 Frankfurt, Germany; 60. Center for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, 2300 Copenhagen, Denmark; 61. Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 62. Department of Public Health, Ben-Gurion University of the Negev, Beer-Sheva, 841050, Israel; 63. National Autism Research Center of Israel, Ben-Gurion University of the Negev, 8410501, Beer-Sheva, Israel; 64. Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114, USA; 65. Department of Child and Adolescent Psychiatry, Hospital General Universitario Gregorio Marañón, IiSGM, CIBERSAM, School of Medicine Complutense University, 28009 Madrid, Spain; 66. Department of Child Psychiatry, Tampere University and Tampere University Hospital, 33520 Tampere, Finland; 67. Program in Genetics and Genome Biology, The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON M5G 0A4, Canada; 68. Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, ON M5S, Canada; 69. Harvard Medical School, Boston, MA 02115, USA; 70. Fundación Pública Galega de Medicina Xenómica, Servicio Galego de Saúde (SERGAS), 15706 Santiago de Compostela, Spain; 71. Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA 02115, USA; 72. Medical Genomics Center, Nagoya University Hospital, Nagoya 466-8550, Japan; 73. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland; 74. The Azrieli National Center for Autism and Neurodevelopment Research, Ben-Gurion University of the Negev, 8410501, Beer-Sheva, Israel; 75. Pre-School Psychiatry Unit, Soroka University Medical Center, 8457108, Beer Sheva, Israel; 76. Children's Hospital of Philadelphia, Philadelphia, PA 19104; 77. Department of Psychiatry, University of Utah, Salt Lake City, UT 84108, USA; 78. Life Span Institute and Kansas Center for Autism Research and Training, University of Kansas, Lawrence, KS 66045; 79. Institute for Glyco-core Research (iGCORE), Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan; 80. Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; 81. Norwegian Institute of Public Health, Oslo, 0213, Norway; 82. Department of Psychiatry, University of Cincinnati, Cincinnati OH 45219 USA; 83. School of Biosystem and Biomedical Science, College of Health Science, Korea University, Seoul 02841, Republic of Korea; 84. Division of Psychological Medicine & Clinical Neurosciences, MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University, Cardiff CF24 4HQ, United Kingdom; 85. Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey; 86. Child Neuropsychiatry Department, Department of Surgical Sciences, Dentistry, Gynecology and Pediatrics. University of Verona, 37134 Verona, Italy; 87. Department of Developmental Medicine, Boston Children's Hospital, Boston, MA 02115, USA; 88. Department of Psychiatry, School of Medicine, Trinity College Dublin, Dublin, D02 VR66, Ireland; 89. Center for Autism Research and Treatment and Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90024, USA; 90. Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA; 91. Department of Biostatistics, Columbia University, New York, NY 10032, USA; 92. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.; 93. Department of Molecular Biology, Cell Biology and Biochemistry and Department of

Psychiatry and Human Behavior, Brown University, Providence, RI 02912, USA; 94. UCB Pharma, 1420 braine-l'alleud, Belgium; 95. Maine Medical Center and Maine Medical Center Research Institute, Portland, ME 04074, USA; 96. Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA; 97. Center for Genomic Medicine, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; 98. Center for Medical Genetics Dr. Jacinto Magalhães, National Health Institute Dr. Ricardo Jorge, 4000-053 Porto, Portugal; 99. New York State Psychiatric Institute, New York, NY 10032, USA; 100. Department of Epidemiology, Mailman School of Public Health, Columbia University, Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA; 101. Department of Psychiatry and Behavioural Neurosciences, Offord Centre for Child Studies, McMaster University, Hamilton, ON L8P 0A1, Canada; 102. National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA; 103. Howard Hughes Medical Institute and Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA 02115, USA; 104. The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; 105. and Departments of Neurology and Pediatrics, Harvard Medical School, Boston, MA 02115, USA; 106. Department of Statistics and Data Science, University of Pennsylvania, Philadelphia PA 19104, USA; 107. Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA; 108. Institute for Neurodegenerative Diseases, UCSF Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA 94143, USA; 109. Quantitative Biosciences Institute (QBI), University of California, San Francisco, San Francisco, CA 94143, USA; 110. Division of Genetics, Boston Children's Hospital, Boston, MA 02115, USA; 111. National Centre for Register-Based Research, Aarhus University, 8210 Aarhus, Denmark; 112. Centre for Integrated Register-based Research, Aarhus University, 8210 Aarhus, Denmark; 113. Department of Biomedicine, Aarhus University, 8000 Aarhus, Denmark.; 114. Centre for integrative Sequencing (iSEQ), Aarhus University, Aarhus, Denmark; 115. Danish Centre for Neonatal Screening, Department for Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark; 116. Institute of Biological Psychiatry, MHC Sct. Hans, Mental Health Services Copenhagen, 4000 Roskilde, Denmark; 117. Genomics plc, Oxford, UK; 118. MRC Centre for Neuropsychiatric Genetics & Genomics, School of Medicine, Cardiff University, Cardiff, UK; 119. Department of Psychiatry, Psychosomatics and Psychotherapy, Center of Mental Health, University Hospital Würzburg, Würzburg, Germany; 120. Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden; 121. Psychosis Research Unit, Aarhus University Hospital, 8240 Risskov, Denmark; 122. Mental Health Services in the Capital Region of Denmark, Mental Health Center Copenhagen, University of Copenhagen, 2200 Copenhagen, Denmark; 123. Department of Psychiatry and Psychotherapy, Universitätsmedizin Campus Charité Mitte, Berlin, Germany; 124. Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, 92093, USA; 125. Department of Clinical Medicine, University of Copenhagen, 2200 Copenhagen, Denmark; 126. Department of Epidemiology, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Building II 2nd Floor, Boston, MA 02115, USA; 127. Metabolomics Platform,Broad Institute of Harvard and MIT, 415 Main St, Cambridge, MA 02142, USA; 128. Verve Therapeutics, Cambridge, MA 02139, USA

SNV/indel processing

ASC+SSC

The de-duplicated set of 24,931 samples of the ASC+SSC cohort was processed identically across two datasets (ASC B14 and ASC B15-16), with raw sequencing outputs aligned to the GRCh38 reference genome and variants jointly called using GATK[66]. Briefly, samples were called individually using local realignment by GATK HaplotypeCaller in gVCF mode, such that every position in the genome is assigned likelihoods for discovered variants or for the reference. Individual gVCF outputs were then post-processed, assigning "blocks" of homozygous reference calls one of three qualities: no coverage/evidence, genotype quality < Q20, or genotype quality ≥ Q20. These compressed per-sample gVCF genotype data, alleles, and sequence-based annotations were then merged using GenomicsDB (https://github.com/Intel-HLS/GenomicsDB), a datastore designed for genomics that takes advantage of a cohort's sparsity of variant genotypes. Samples were jointly genotyped for high confidence alleles using GenotypeGVCFs on all autosomes. Variant call accuracy was estimated using Variant Quality Score Recalibration (VQSR) as in the GATK Best Practices.

SPARK

For the SPARK Pilot, we accessed GRCh38-aligned bam files for 1,379 samples from SFARI via the /SPARK/Pilot/BAM/ directory in the Globus file sharing platform. The bam files were subsequently realigned to the GRCh38 reference and fed into the same GATK pipeline as outlined above for the ASC+SSC data. For the SPARK main freeze, 27,270 individual GATK-produced gVCFs were downloaded from SFARI via the /SPARK/Regeneron/SPARK_Freeze_20190912/Variants/GATK/ directory in Globus. Per provided documentation, the gVCFs were generated with GATK v4.1.2.0 HaplotypeCaller with default thresholds for calling, using the GRCh38 reference and target files provided by Regeneron (genome.hg38rg.fa & xgen_plus_spikein.b38.bed, respectively). New quality scores, lenient processing of VCF files, and 100bp padding for intervals were also used. We carried out subsequent joint calling of all 27,270 sample gVCFs via GATK to produce one unified VCF.

Previously published variants

De novo variants from 559 samples (458 probands and 101 siblings; counts include samples with no variant) were directly carried over from Satterstrom *et al.*; they had been curated from a number of previously published studies[45,64,72–74]. The genes that each variant impacted as well as the type of impact (PTV, missense, etc) were retained.

SNV/indel filtering

Creation of working datasets

VCF processing was carried out in Hail 0.2 (https://hail.is) to create a working dataset for each of the four batches of samples. After VCF import, multi-allelic sites were split and variants were annotated using the Variant Effect Predictor (VEP)[75]. Hail's ibd() function was used to verify reported pedigrees and check for duplicate samples within and across datasets. Annotations such as gene and functional consequence were assigned to each variant by prioritizing coding canonical transcripts in the VEP output. Low-complexity regions (using https://github.com/lh3/varcmp/blob/master/scripts/LCR-hs38.bed.gz) were removed. Sexes were imputed with the impute_sex() function in Hail, after which several genotype filters were applied.

Genotypes were filtered to remove calls with a depth below 10 (except for male hemizygous regions, where calls were removed if depth was below 7) or above 1000. Homozygous reference calls were filtered if the genotype quality (GQ) was below 25, while heterozygous and homozygous variant calls were filtered if the phred-scaled likelihood of the call being homozygous reference (PL[HomRef]) was below 25. Additionally, heterozygous calls were dropped if they were in a hemizygous region in a male sample, and any call apparently on the Y chromosome in a female sample was dropped.

Genotypes were further filtered if the allele balance (that is, the number of reads supporting the alternate allele divided by the depth) of a heterozygous call was below 0.25; if the probability of the allele balance (based on a binomial distribution with mean 0.5) was below 1e-9; or if the number of informative reads supporting a heterozygous call (counting reads supporting either the reference or alternate allele) or homozygous alternate call (counting reads supporting the alternate allele) was less than 90% of the depth. Variants were then dropped if they had a call rate below 10% or a Hardy-Weinberg p value less than 10^{-12}, and a working dataset was written.

De novo variant calling and quality control

To call *de novo* variants from the working datasets, a filter requiring a GQ of at least 25 was applied to every genotype, and Hail's de_novo() function was called with variant frequencies from the non-neuro subset of gnomAD GRCh38 exomes v2.1.1 (gs://gnomad-public/release/2.1.1/liftover_grch38/ht/exomes/gnomad.exomes.r2.1.1.sites.liftover_grch38.ht) used as priors. After calling, putative *de novo* variants were dropped if they were present within this gnomAD subset at a frequency greater than 0.1%, or if they were present within their own dataset at a frequency greater than 0.1%. Variants were also dropped if they contained "ExcessHet" in the Filters field, had a proband allele balance of less than 0.3, or had a depth ratio (child read depth divided by the sum of parental read depth) of less than 0.3.

The following filtering steps were dataset-specific. For the ASC B14 dataset, variants were kept only if the calling algorithm marked them as "HIGH" or "MEDIUM" confidence, with the medium-confidence calls limited to a maximum allele count in the dataset of 3. SNPs with a VQSLOD below -20 were dropped, as were indels with a VQSLOD below -2. The proband allele balance threshold was raised to 0.4 for variants from cell line-derived samples, and variants that appeared more than three times in the remaining set were dropped (this dropped two different synonymous variants in *OR13C2*). For the ASC B15-B16 dataset, only "HIGH" or "MEDIUM" confidence variants with a VQSLOD at least -11.6 were kept, with the medium-confidence calls limited to a maximum allele count in the dataset of 1 and the call rate threshold raised to 0.9. For the SPARK Pilot dataset, only "HIGH" confidence variants were kept. SNPs were dropped if they fell in a VQSR tranche above 99.5, and indels were dropped if they fell in a VQSR tranche above 96.0. As with the B15-B16 dataset, the call rate threshold was raised to 0.9. One call was dropped that had reads supporting the alternate allele in a parent's homozygous reference genotype. For the SPARK main freeze, only "HIGH" or "MEDIUM" confidence variants were kept, with the medium-confidence calls limited to a maximum allele count in the dataset of 3. Variants with a VQSLOD below -11.6 were dropped. The call rate threshold was set at 0.5, and variants that appeared more than three times in the remaining set were dropped (this dropped one intronic variant in *THBS2* and another in *HLA-A*).

Following these dataset-specific steps, one variant was selected per person per gene, prioritizing variants with more severe consequences. Then, for each dataset, samples were dropped if their count of coding *de novo* variants was significantly greater (i.e. $p < 0.05/n$) than expected based on a Poisson distribution with the dataset's observed mean number of coding

variants per sample (in ASC B14, this dropped 28 samples with more than 9 coding *de novo* variants; in both ASC B15-16 and the SPARK Pilot, this would have dropped any samples with more than 7, but there were none; in the SPARK main freeze, this dropped 3 samples with more than 9). Finally, indels were dropped if the probability of their allele balance (based on a binomial distribution centered around 0.5) was less than 0.05.

Transmitted variants

For evaluation of transmitted/untransmitted variants, dataset-specific filters were applied: variants were required to possess a VQSLOD value of at least 5.13 in the ASC B14 dataset, 24.12 in the ASC B15-B16 dataset, and 6.01 in the SPARK main freeze. In the SPARK Pilot, all variants that did not pass VQSR were dropped. These thresholds were based on identifying the VQSLOD values necessary to balance the transmission/non-transmission of singleton synonymous SNVs from the parents or the transmission/non-transmission of non-coding indels, and selecting the more stringent VQSLOD value. For the ASC B14 dataset in particular, additional filters were applied to standardize data from samples originally sequenced across a span of several years. Specifically, variants were retained only if they possessed a strand odds ratio no greater than 3, a Read Position Rank Sum of at least -8, and a quality by depth of at least 1 (for SNVs) or 3 (for indels). After application of these filters, final counts of transmitted and non-transmitted alleles were produced for variants of different classes (e.g., PTV, MisB, MisA).

CNV processing

Cluster curation by PCA analysis

A set of 7,981 target regions were chosen from among 7 popular exome enrichment kits (Agilent_V4, Agilent_V5, Agilent_V6r2, Agilent_V7, NimbleGen SeqCap EZ2, NimbleGen SeqCap EZ3, Illumina TruSeq) such that each chosen region was uniquely targeted by 1 of the 7 kits. Depth of coverage was collected over the 7,981 regions for all samples, and a primary PCA analysis was conducted to identify samples that were prepared with different enrichment kits (**Supplementary Fig. 1**). After kit determination, a secondary PCA analysis was conducted for samples on the same enrichment kits to determine final cluster assignment via a hierarchical clustering approach.

Curation of intervals and exons modeled

Due to the differential exon targeting and differences in transcriptome annotation, we standardized the set of intervals over which GATK-gCNV queried for CNV events to better homogenize our cohort. To construct this list of intervals, we took all exons from the Gencode V33 annotation and collapsed them to non-overlapping intervals. Collapsed intervals longer than 800bp were evenly split into subdivisions less than 1,600bp, in an attempt to increase the number of intervals/data points for GATK-gCNV to identify copy number events. The resulting intervals were padded by 100 basepairs in a non-overlapping, equal manner. Intervals which had <50% of samples with <10 reads, >50% segmental duplication content, <90% mappability, or GC content outside of the interval (0.1, 0.9) were excluded from further processing.

Cluster processing with GATK-gCNV

GATK-gCNV can be run in 2 modes: cohort or case. Cohort mode takes the set of input samples, constructs a model on the copy number states of each input region that passes filtering metrics, and subsequently makes determinations of the copy number events in those samples. Case mode takes a pre-computed model output from cohort mode and proceeds to directly determine copy number events in a set of supplied samples, offering time and cost savings if the samples are similar enough. Cohort mode was run for every cluster determined by PCA analysis, using 200 samples per cluster. As most PCA-defined clusters consisted of more than the 200 samples used to create the model in GATK-gCNV cohort mode, the remaining samples from each cohort were analyzed with GATK-gCNV using case mode with their respective cohort mode models.

GATK-gCNV parameters included:
num_interval_scatter = 12500
gcnv_interval_psi_scale = 0.01
gcnv_max_bias_factors = 6
gcnv_p_active = 0.1
gcnv_o_alt = 0.0005
gcnv_sample_psi_scale = 0.01

Variant compilation, quality control, and annotation

For each sample processed using GATK-gCNV, we extracted non-reference copy number events from the output. All calls with quality score (QS) ≥ 20 were considered for defragmentation (concatenation). Each candidate call was translated onto the interval-space and extended by 50% in width on both ends; calls that had consistent copy number and overlap when extended were merged into a single call.

All calls for samples and clusters were compiled into a single table. Then single-linkage clustering requiring 80% reciprocal overlap was used to determine when multiple calls were the same site. Site frequencies were assigned to each site as the proportion of parents that had the site in the combined dataset. We collated the number of captured intervals (those passing the genomic content and minimum median read count thresholds) that each variant spans, as well as the number of captured exons that the captured intervals corresponded to when mapped onto canonical transcripts of protein-coding genes in the Gencode V33 annotation.

For quality control of variants, we used sample-, site-, and call-level metrics. Sites were retained if the site was rare (site frequency less than 1%) and spanned more than two captured exons. Individual call-level filters were implemented as a function of the number of well-captured intervals that the variant spanned. For homozygous deletions, QS threshold = min(400, 10*number of intervals); for heterozygous deletions, QS threshold = min(100, 10*number of intervals); and for duplications, QS threshold = min(50, 4*number of intervals). For sample-level quality control, samples were kept if their number of raw, autosomal CNV calls detected by GATK-gCNV did not exceed 200 and if the number of calls with QS≥20 did not exceed 35.

We annotated a deletion as impacting a gene if ≥10% of the non-redundant exon-basepairs of the gene were overlapped by the deletion; for a duplication, a gene was impacted if ≥75% of the non-redundant exon-basepairs were overlapped by the duplication. CNVs were also annotated against a list of 79 curated genomic disorder (GD) loci (**Supplementary Table 10**). A call was considered a genomic disorder CNV if it shared 50% reciprocal overlap with the annotated GD.

For GDs that are terminal on a chromosome, a CNV that overlapped at least one of the five genes closest to that terminal was labeled as a terminal GD CNV.

Familial CNV delineation

For samples from a family where all 3 members were characterized for CNVs, we further determined transmitted and *de novo* CNVs. High quality CNVs in offspring that had less than 30% of its intervals or genomic coordinates covered by a matching, unfiltered call in either parent were deemed candidate *de novo* CNVs and underwent secondary screening. They were considered *de novo* only if the following criteria were met: average normalized copy number was within 0.3 of the reported copy number; the minimum normalized copy difference between the child and either parent was > 0.7; and less than 50% of the normalized copy number of the constituent intervals had median absolute deviation across other samples greater than 0.5 (*i.e.* the intervals were well-normalized). Variants were considered transmitted from a mosaic parent if the normalized copy number of the parents across that site showed skew towards the offspring's copy number, using a threshold determined as a function of the number of intervals spanned by the CNV. Conversely, high-quality rare CNVs in the parents for which more than 30% of the captured intervals that overlapped unfiltered CNVs in the offspring were deemed inherited CNVs.

Parent-of-origin determination

De novo CNVs were evaluated for parent of origin where possible using the SNP/indel joint-called VCF. For each *de novo* deletion in the child, the region was queried for variants in the offspring and both parents. For each variant in the child within that region that was either 0/0 or 1/1, we recorded which parent had the corresponding variant and which had the reference. For *de novo* duplications, we recorded variants for which the child was 0/1 and had alternate-allele frequency > 60% or < 40%. For those variants, we then recorded if one parent was homozygous reference while the other parent was not. A binomial test of the number of variants that are consistent with the mother versus the father was carried out for binomial parameter p=0.5. Parental origin was called for sites where the binomial test p-value was less than 0.125.

CNV filtering for association modeling

To reduce false positive signals that are detrimental to association study of rare variants, we implemented additional CNV filtering. For rare and *de novo* CNVs, we required the frequency of non-diploid events of a particular site to be less than 1% in addition to the site frequency being less than 1%. This filtering was carried out to remove the limited occurrences where a site harbors a common deletion and a rare duplication, or vice versa, for which exome CNV delineation remains challenging. We also required that less than 60% of the intervals of a site harbor evidence of small, cluster-specific common CNVs (those spanning <20 intervals and with >2.5% cluster-specific frequency). For *de novo* variants, we further increased the minimum QS threshold to 200 regardless of the size of the event.

CNV Odds Ratio Comparison to UK Biobank

We compared GD CNV rates in 143,532 UK Biobank exome sequencing samples which we know to be unaffected by neuropsychiatric or developmental phenotypes to the 13,786 offspring with ASD (which included *de novo* and transmitted events). Both datasets were processed with identical implementations of the GATK-gCNV pipeline (Babadi *et al.* in preparation). To calculate

the OR of ASD risk for each GD locus, we applied Fisher's exact test to the number of carriers versus non-carriers detected in the ASD dataset compared to the number of carriers versus non-carriers in the UK Biobank dataset. We found a strong inverse relationship between frequency of GDs in the UK Biobank and risk of ASD for carriers of GD events.

Relative Difference Calculation

Due to the occurrence of multiple observations of transmitted MisA variants in the same individual, the calculation of an odds ratio would require binary classification of an individual to having or not having a variant, resulting in information loss. To avoid this and in order to present a comparable measurement of enrichment signal between *de novo,* inherited, and case/control variation across all variant types, we calculated a relative difference statistic and its associated confidence intervals. The relative difference statistic is: (a) for *de novo* variants, the average rate of variants in affected offspring divided by the average rate of variants in unaffected offspring; (b) for inherited variants, the average rate of transmitted variants divided by the average rate of untransmitted variants in affected offspring; (c) for case/control, the average rate of variants in cases divided by the average rate of variants in controls from case/control data. The confidence interval of these statistics is calculated based on a binomial test, with probability of the binomial distribution equal to (a) for *de novo*, number of affected offspring divided by the total number of offspring; (b) for inherited, 0.5; and (c) for case/control, the number of cases divided by the total number of samples in the case/control data.

TADA Bayesian Framework for Gene Association

Modes of evidence included

We have expanded evidence integration to all of the following combinations, comprising a full 5x3 combination of variant types by inheritance classes (PTV, MisB, MisA, deletion, duplication) x (*de novo*, case-control, inherited).

Preparing CNVs for TADA Integration

Due to the challenge in pinpointing driver genes within large, recurrent CNVs, we opted to exclude NAHR-mediated CNVs and focus on modeling smaller CNVs that impact 8 or fewer constrained genes (defined by LOEUF < 0.6).

Prior Elicitation for TADA using Empirical Bayes

PTVs

We used an empirical Bayes approach, borrowing information across genes, to estimate the prior relative risk (gamma) for ASD association of each gene. This consists of combining 2 components: (1) the relative risk of PTVs for each inheritance type and (2) the estimated fraction of genes that substantially influence ASD risk. Gamma for each gene is then calculated as (1) divided by (2), smoothed over a measure of constraint. In Satterstrom *et al. Cell* 2020, the smoothing is conducted over pLI, which is now superseded by the updated loss-of-function observed/expected upper bound fraction (LOEUF)[34]. In Satterstrom *et al.*[4], the fraction of genes that substantially influence ASD risk is assumed to be constant over the entire pLI range. We relaxed this assumption because ASD-associated genes identified to date are heavily concentrated among constrained rather than unconstrained genes. We used equations in the

supplement of He *et al.*[36] to estimate the proportion of genes that are risk genes, specifically methods of moments estimation for the proportion of risk genes:

$$(\bar{\gamma} - 1)k = \frac{C}{2N\bar{\mu}} - m. \quad (29)$$

$$k\mathrm{E}\left(p|\bar{\gamma}, \beta\right) + (m - k)\mathrm{E}\left(p|\gamma = 1\right) = M, \quad (33)$$

C = the observed number of PTV events in all genes
M = the number of genes with more than 1 PTV event
m = the number of genes
N = the number of probands
k = the number of ASD genes
γ= the relative risk

(33) is [number of multi-hit genes under alt] + [number of multi-hit genes under risk=1]

Using equations (29) and (33) and by using a sliding window of 20% of the genes, ordered by LOEUF, we computed a rolling-average estimate for the fraction of risk genes (**Supplementary Fig. 3**).

For the *de novo* prior, (1) is estimated as the relative enrichment of variants in affected versus unaffected offspring; for the case-control prior, (1) is estimated as the relative enrichment of variants in cases versus controls; and for the prior of inherited variants, (1) is estimated as the relative enrichment of transmitted versus untransmitted variants. As expected, the estimated prior relative risks all decrease with lower genic constraint, with the *de novo* prior being the strongest, followed by case-control and then inherited, respectively. We averaged the priors between our ASC/SSC cohort and the SPARK cohort.

Missense variants
In Satterstrom *et al.*[4], the prior relative risks of *de novo* MisA and MisB variants were separately estimated as (R-1)/0.05+1, where R is the enrichment ratio of the variant class, and 0.05 is the proportion of risk genes across the genome. We adopted the same approach. For *de novo*, R is the ratio of observed counts from affected versus unaffected offspring; for case-control, R is the ratio of observed counts from cases versus controls; and for inherited, R is the ratio of observed counts of transmitted to untransmitted variants from parents to their affected offspring. All priors are floored at 1.05x relative risk.

CNVs
To estimate the prior relative risk of *de novo* deletion CNVs, we use the estimated prior risk for *de novo* PTVs falling in constrained genes within the CNV. Specifically, the estimated prior risk of a CNV is the summation of the estimated PTV prior risk for constrained genes that are annotated for that CNV. For *de novo* duplications, we first computed the same sum, then downweighted the sum proportional to the observed ratio of *de novo* deletions/duplications. This accounted for the decreased penetrance of duplications compared to deletions. For inherited and case/control CNVs, we followed the same approach, substituting the corresponding prior PTV risk estimate.

BF calculation using TADA

De Novo - PTV/MisA/MisB

For each gene, PTV, MisA, and MisB BFs were calculated per a previously described implementation of TADA[4,36], taking into account sample size and using updated mutation rates and prior relative risk. Mutation rates for PTVs were updated using gnomAD (v2.1.1) LOF mutation rates, scaled so that the expected number of *de novo* PTVs matched the observed number in unaffected siblings. MisA and MisB mutation rates were carried over from Satterstrom *et al.*, and genes previously missing MisA or MisB mutation rates were given estimates using gnomAD (v.2.1.1) missense mutation rates scaled by the average missense to MisA or MisB mutation ratio respectively. MisA and MisB mutation rates were scaled so that expected and observed *de novo* counts matched that of unaffected siblings.

BFs for *de novo* PTV/MisA/MisB variant types were calculated separately for our ASC/SSC sub-cohort and the SPARK cohort, to facilitate direct comparisons to *de novo* PTV/MisA/MisB evidence from the DDD study.

De Novo - CNVs

BFs were estimated per CNV, analogously to the per-gene calculation of BFs for PTVs/MisA/MisB variants, and separately for deletions and duplications. We excluded from modeling CNVs that were NAHR-mediated GD events and focused on the contribution of constrained genes within CNV segments. Mutation rates of deletions and duplications were estimated using data from the unaffected siblings, where the rate was approximated as the number of observed *de novo* deletions or duplications that span X constrained genes, divided by the total number of sequences of X constrained genes. Given the prior relative risk and the mutation rate, each unique CNV had a total BF estimate which was then distributed to the constituent constrained genes relative to their LOEUF measure. For example, if a CNV had a total BF=10 and it spanned two constrained genes, A and B, with respective LOEUF values of 0.1 and 0.4, we distributed BF=8 to gene A and BF=2 to gene B, consistent with the 4:1 relative constraint ratio estimated by LOEUF (where lower values signal a more constrained gene).

Case/control - PTV/MisA/MisB

For each gene, PTV, MisA, and MisB BFs were calculated per a previously described implementation of TADA[4,36], taking into account case/control sample sizes and updated priors for case/control variants.

Case/control - CNV

Like *de novo* CNVs, BFs were estimated per CNV using the case/control PTV/MisA/MisB framework described previously, with prior relative risk estimated as the sum of the prior case/control PTV relative risk of the constituent constrained genes. For duplications in this setting, we also used a downweighting factor for prior relative risk compared to a deletion of the same gene, analogous to the *de novo* CNV setting. The BF for each CNV was then distributed to the genes that constitute the CNV, proportional to LOEUF, as in the *de novo* setting.

Inherited - PTV/MisA/MisB

For each gene, PTV, MisA, and MisB BFs were calculated per a previously described implementation of TADA for case/control data[4,36], using updated priors for inherited variants as described above, and using a sample size of all affected offspring in trio settings.

Inherited - CNV

For inherited CNVs, BFs are estimated per CNV, using the case/control PTV/MisA/MisB framework described previously, with prior relative risk estimated as the sum of the prior inherited PTV relative risk of the constituent constrained genes. Duplications also used a downweighting factor for prior relative risk compared to a deletion of the same gene, analogous to the *de novo* CNV setting. The BF for each CNV was then distributed to the genes that constitute the CNV, proportional to LOEUF.

BF integration using TADA

For each variant class (PTV, MisB, MisA, deletions, and duplications), we set a floor of 1 for the BF, so that evidence against association based on one variant type would not degrade evidence for association from another variant type. For example, in some genes, PTVs dominate the signal for association while missense signal is largely lacking; in others, the pattern is reversed. We believe such patterns stem from how the allelic architecture relates to risk for ASD, such that setting a floor for the BF preserves signals related to allelic architecture.

Within each variant class, however, we allowed the BF evidence to offset each other when appropriate. For instance, BF > 1 in unaffected siblings, arising from non-zero *de novo* counts in siblings, are used to offset the BF of that same gene in affected probands. Operationally, this is achieved by calculating the BF for a certain gene contributed by PTVs as: BF_PTV = floor(BF_PTV_probands / BF_PTV_siblings, 1).

Finally, the total gene-level BF was calculated as the product of the BFs across the 5 variant classes: BF = BF_PTV * BF_MisA * BF_MisB * BF_deletion * BF_duplication. Of note, to ensure no gene was nominated from CNV evidence alone, we set BF_deletion and BF_duplication to 1 if the total BF evidence of PTV/MisA/MisB did not exceed 5.

All BFs are subsequently converted to a posterior probability, which is in turn used to compute a FDR to nominate genes of relevance.

Cross-validation relation of FDR with relative risk

To relate the estimated FDR to a relative risk measure of *de novo* variants, we conducted cross-validation. We randomly subset the data into 10 equal shards, taking turns to estimate the TADA model using 9 folds while holding the remaining shard out. For each model fit, we calculated the FDR of each gene and examined the empirical relative risk of variant types and inheritance classes between affected and unaffected individuals for differing FDR thresholds. We repeated this process 1,000 times, randomly assigning samples to the 10 shards during each iteration. We found cross-validated relative difference estimates of 18.5, 10, and 3.5 for de novo PTVs, MisB variants, and MisA variants, respectively (**Supplementary Table 12**).

Female protective effect versus ascertainment bias of affected females

We address these hypotheses by assessing a measure of ASD severity, namely the calibrated severity score for the ADOS (ADOS.CSS), and a measure of intellectual functioning, IQ. We have 2,095 individuals who have a score on the ADOS.CSS (**Supplementary Table 13a**) and are characterized for both *de novo* SNVs and CNVs. Of these individuals, 268 are females, 1,827 are males, and 104 of them carry a *de novo* SNV (PTV or MisB), or a *de novo* CNV overlapping one of the 185 TADA genes, or a *de novo* CNV in a GD locus (Carrier; **Supplementary Table 13a**). Following Chaste *et al.*[76], we partition ADOS.CSS into two groups < 8 and ≥ 8, which correspond to those who are less severely impaired and more severely

impaired, or moderate and severe ASD. This split yields 1,075 individuals who are moderate ASD and 1,020 who are severe ASD (**Supplementary Table 13a**).

These data were analyzed using logistic regression with carrier status as the outcome and sex, severity, and their interaction as predictors. In this model, the interaction has no significant effect (Odds Ratio OR: 0.72, p = 0.46), and neither did the main effect (Sex: Odds Ratio OR: 1.8, p = 0.083; Severity: Odds Ratio OR: 1.8, p = 0.14). Because the interaction was non-significant, we re-analyzed the data using only main effects. In the simpler model, sex (OR: 2.2; $p=2.7x10^{-4}$) was significant, while the level of severity approached significance (OR: 1.4; p=0.069). Females were at roughly two-fold higher risk for being carriers than males, while individuals with a more severe ASD phenotype were more likely to be carriers, although not significantly so.

Next, we turned to IQ as a measure of severity. Again, we split the data according to IQ threshold: we took < 70 to be low functioning, with values above 70 considered high functioning. For SPARK data, we used the "reported_cog_test_score" field; for ASC/SSC data, we used full-scale IQ. There were 5,460 individuals for whom IQ was reported and data were complete for mutation status, 1,503 of whom were low functioning and 3,957 of whom were high functioning (**Supplementary Table 13b**), and with similar sex (4,610 males, 850 females) and carrier status (367 Carriers, 5,093 Non-carriers) distributions as in **Supplementary Table 13a**.

These data were analyzed using logistic regression with carrier status as the outcome and sex, level of function, and their interaction as predictors. There was a significant effect for sex (Odds Ratio OR: 2.1, p = $p=9.5x10^{-5}$), as females were roughly twice as likely to be carriers than males. Level of function also had a significant effect (OR: 2.5, p = $p=5.8x10^{-5}$), with low-functioning individuals more likely to be mutation carriers, while the interaction was not significant (OR: 0.72, p = 0.21). Next, we fitted the data using only main effects. In the simpler model, both sex (OR: 1.8; $p=1.2x10^{-5}$) and level of function (OR: 1.9; $p=2.3x10^{-9}$) were significant. For both the interaction and main effects models, females were at roughly a two-fold higher risk for being carriers than males.

In both of these analyses, we observed no evidence that the increased burden in females was due to an ascertainment bias, which would manifest as an interaction effect. Instead, sex and severity, however the latter is defined, appeared to combine additively to determine liability. Thus, we concluded that the increased *de novo* mutation burden in females, relative to males, was largely due to the female protective effect.

We next asked if the expanded list of NDD genes might alter this conclusion. To evaluate this possibility, we used the list of 373 NDD genes and GD loci to define carrier status. Fitting data for either ADOS.CSS (**Supplementary Table 13c**) or IQ (**Supplementary Table 13d**), we reached identical conclusions. The interaction was not significant for the data in either table. For the main effects model, females were roughly two-fold more likely to be carriers than males (OR:2.5, $p=3.6x10^{-6}$ for ADOS.CSS and OR:2.0, $p=2.6x10^{-9}$ for IQ). For these data, severity of ASD or mental function was significant and of effect size similar to that found previously (OR:1.6, p=0.008 for ADOS.CSS and OR:2.0, $p=1.0x10^{-10}$ for IQ).

Tree analysis

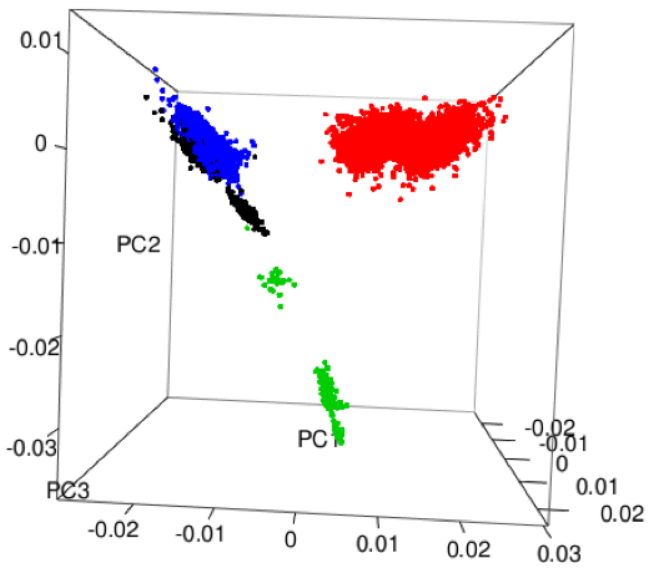
Applying cFIT to fetal cells from two studies, which we will call “Nowakowski”[55] and “Polioudakis”[58], we obtained integrated factor loadings and gene expression for all measured cells (**Supplementary Fig. 7**). We started with the preprocessed data from those two studies

where the QC had already been done. For more details of the analysis, see Peng *et al*.[56]. The algorithm was applied to a subset of informative genes, including 3,790 genes selected using the Seurat function FindMarkers[77]. To ensure we included all informative ASD and DD risk genes, we also included 676 for which the TADA analysis yielded a q-value < 0.05 for at least one of the three cohorts: ASD, DD, and ASD+DD. After removing genes expressed in less than 5% of cells in either data set, 454 risk genes remained. In total, we identified 3938 genes, which are the only genes included in the downstream analyses.

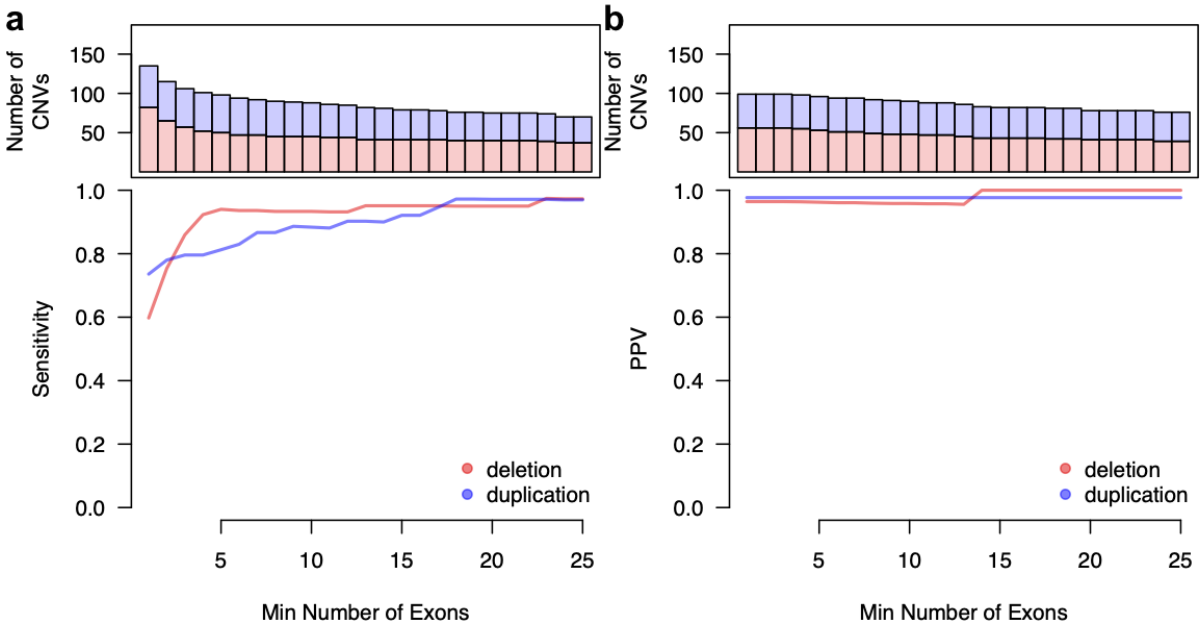
With the integrated data, we applied unsupervised clustering to identify cell subtypes. While it is typically impossible to definitively determine the right number of clusters, it is clear that cell types should follow a hierarchical structure: major clusters should develop into minor more specialized subtypes. Our MRtree method[57] was developed based on this idea (MRtree constructs a hierarchical cluster tree from multiresolution clustering to recover levels of specification among cell clusters; the method relies on a stability measure to cut the tree to obtain stable clustering). To obtain the flat clustering required by MRtree, we used Seurat v4, with resolution varying from 0.05 to 2. By further increasing the resolution, functionally more specialized cell types were separated. After applying the stability criterion to determine the finest stable resolution, the algorithm produced 21 terminal clusters and most exhibited a good mix of cells from each source (**Supplementary Fig. 8a**). The exceptions were two clusters (newborn IN and MGE RG/IPC) derived entirely from Nowakowski cells, which were from a region of the brain not sampled in the Polioudakis study, and one small cluster labeled as ExN, which derived entirely from Polioudakis, which appears to be problematic based on the UMAP (**Supplementary Fig. 8b**). This cluster was removed from downstream analyses. To avoid the tree growing too large, we also excluded undefined types and well-separated clusters from tree construction, including 939 cells with missing labels or labeled as U1, U2, U3, U4, Choroid, End, Endothelial, Per, Glyc, Microglia, Mic, and Mural from two studies.

Functional Description of DD- and ASD-predominant enriched cell types

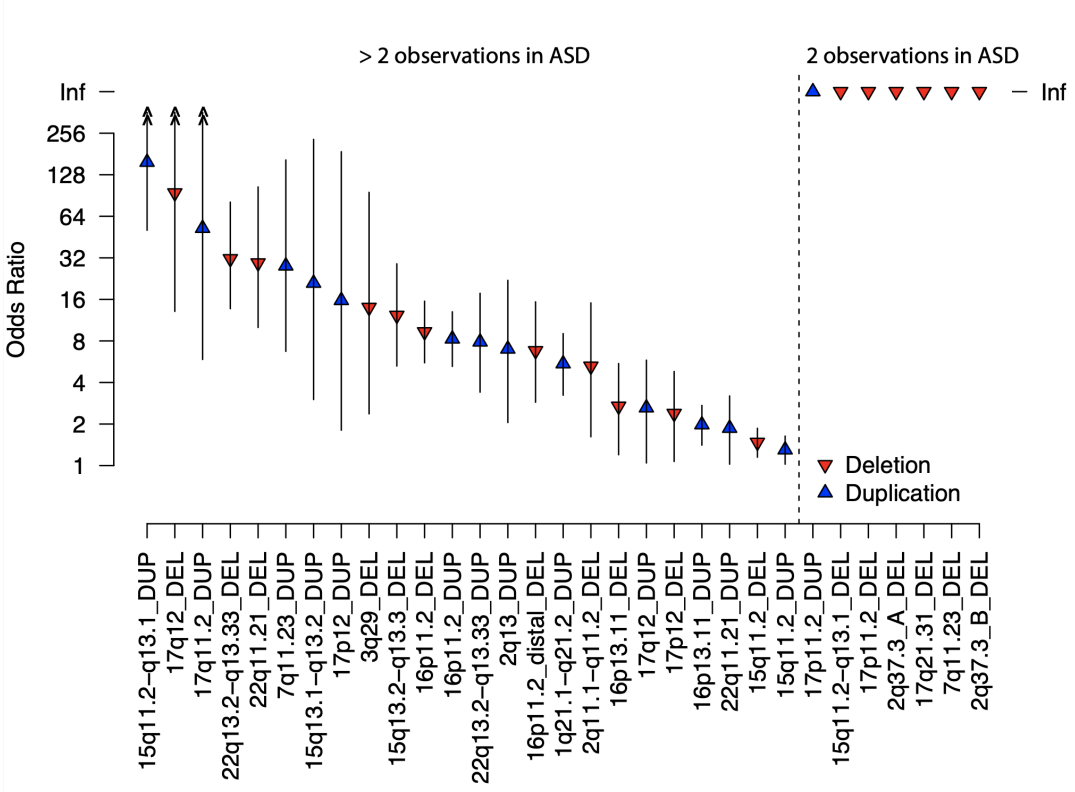
The migrating excitatory neuron cluster (ExN3) consists of newborn neurons beginning migration and differentiation in the dorsal telencephalon. Expressed DD-predominant genes in this cell type include cell-type specific marker *BCL11B* (*CTIP2*[78]) and genes that determine migration (*KIF1A*, *RAC1*[79,80], *SET*, *ZMIZ1*[81,82] and *NFIA*[83,84]). In maturing excitatory neurons (ExM2), DD-predominant genes are related to synaptic membrane dynamics and signaling – *SNAP25*, *KRAS, and PIK3R1* illustrate a move away from migration and an early stage of developing synaptic connections. The DD-predominant genes expressed in intermediate progenitors (IP) and caudal ganglionic eminence (InCGE) cells also highlight their identity and state: both express *SOX2*, *FOXG1*[85], *SMARCA4*, and *ZEB2*[86], all transcription factors important for proliferation and neuronal development. Developmentally, interneurons are expected to be behind maturing excitatory neurons, which explains why InCGE and IP DD-predominant genes are more similar. Interestingly, *ARID1B*[87] is uniquely enriched in InCGE interneurons; both *ARID1B* and *ZEB2*[88] support differentiation of interneurons. By contrast, *CTNNB1*'s unique expression in the IP cluster illustrates its role in regulating proliferation/population size through WNT signaling[89]. Regarding ExMU1, *ASH1L* was recently shown to be important for superficial layer neuronal development[90], and *RORB* is a marker for Layer 3-4 Excitatory Neurons[91]. *ANK2*[92], *TAOK1*[93], and *SCN2A*[94] are all necessary for dendritic complexity. *NRXN1*[95] and *DPYSL2*[96] together support axonal guidance, endocytosis, and synaptogenesis.



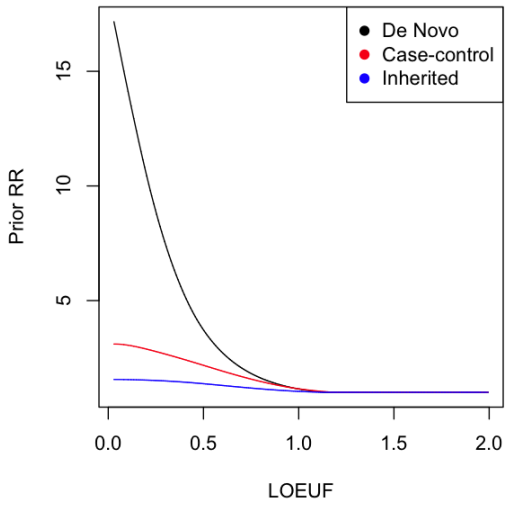
Supplementary Fig. 1. *PCA of read counts from exome sequencing data of 7,981 specially chosen regions. Clear capture kit and technical artifacts can be observed. The first 3 PCs of the chosen target regions clearly differentiate the capture kits used to enrich each sample. Red corresponds to Illumina, black to Agilent, blue to Agilent 1.1, and green to Nimblegen. Secondary clustering in 3d space of each of the determined capture clusters was carried out to create batches for GATK-gCNV to process for CNVs.*



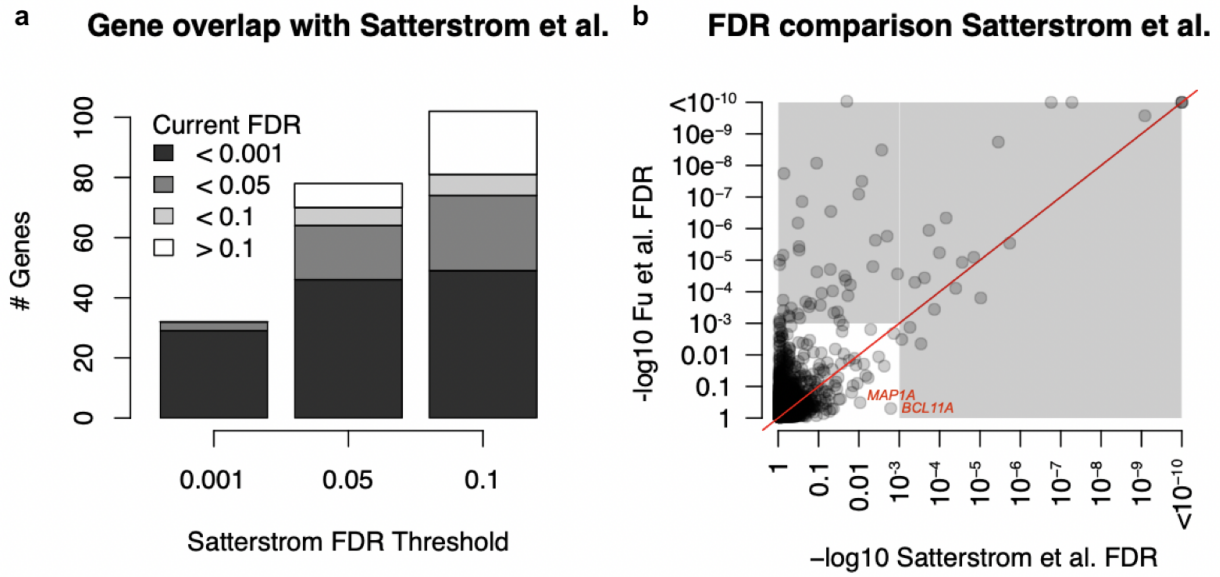
Supplementary Fig. 2. *Using a set of 3,097 trios for which we have matching WES data and gold-standard WGS CNV calls[47], we benchmarked the performance of de novo CNVs derived from WES data using GATK-gCNV as a function of the number of captured exons of canonical transcripts.* ***a****, Sensitivity: using de novo WGS CNVs as ground truth, we find that GATK-gCNV achieves 83% sensitivity at recapitulating gold-standard WGS calls spanning more than two captured exons.* ***b,*** *PPV: given de novo WES CNVs, we looked for corroborating calls from the gold-standard WGS callset. We find that GATK-gCNV achieves 97% PPV for de novo CNVs that span more than two exons.*



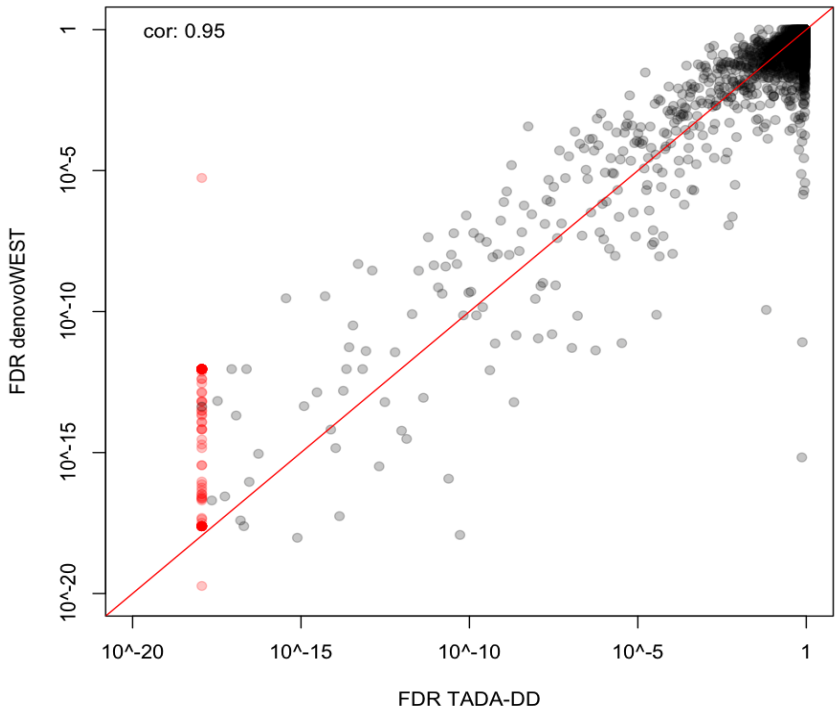
Supplementary Fig. 3. *Using 143,532 identically processed exome sequenced controls from the UK Biobank as the background, we estimated the increased odds ratio of ASD resulting from GD events in our 13,786 offspring with ASD (nominal enrichment, p<.05 shown here; 95% confidence interval from Fisher's exact method are shown as error bars), number of unique samples are 13,786 cases and 143,532 controls.*



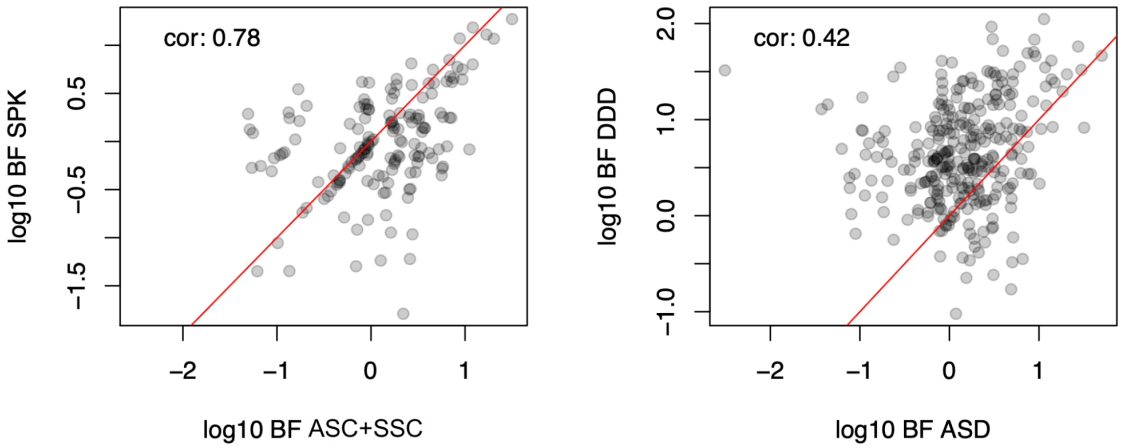
Supplementary Fig. 4. *Based on our prior elicitation procedure, we computed separate priors for the risk of de novo, case/control, and inherited variants in our ASD dataset. Increasing risk trends strongly with increasing constraint, as measured by LOEUF, while de novo variants represent more risk compared to case/control and inherited variants given the same level of constraint.*



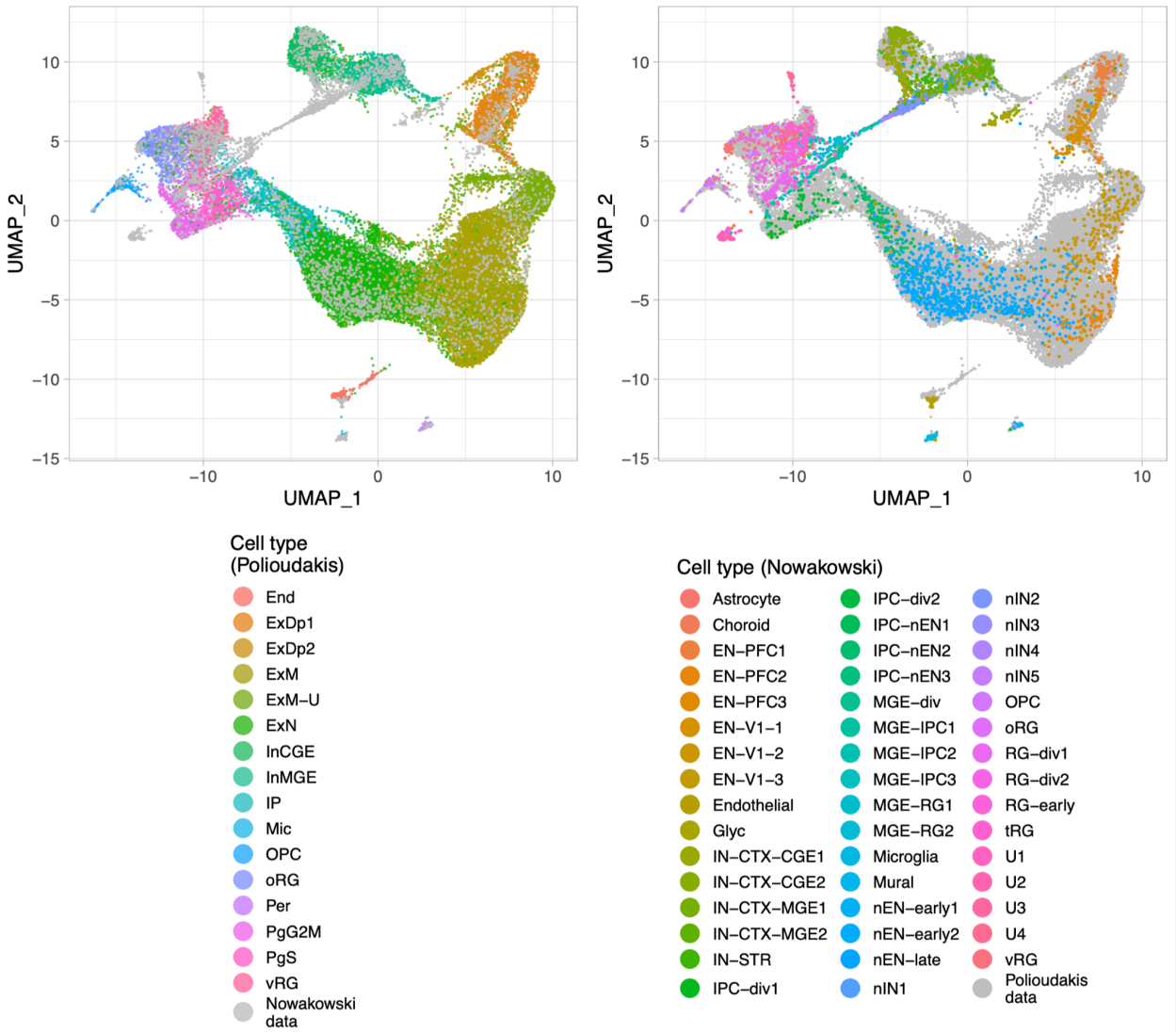
Supplementary Fig. 5. ***a***, *Portrayal of the overlap of Satterstrom et al. 102 genes with the current TADA-ASD FDRs, broken down by the FDR threshold reported in Satterstrom et al.* ***b***, *A comparison of the FDR values for each gene between Satterstrom et al. and the current implementation of TADA for the analysis of ASD risk genes, illustrating the improved power derived from the increased sample sizes and new implementation of TADA.*



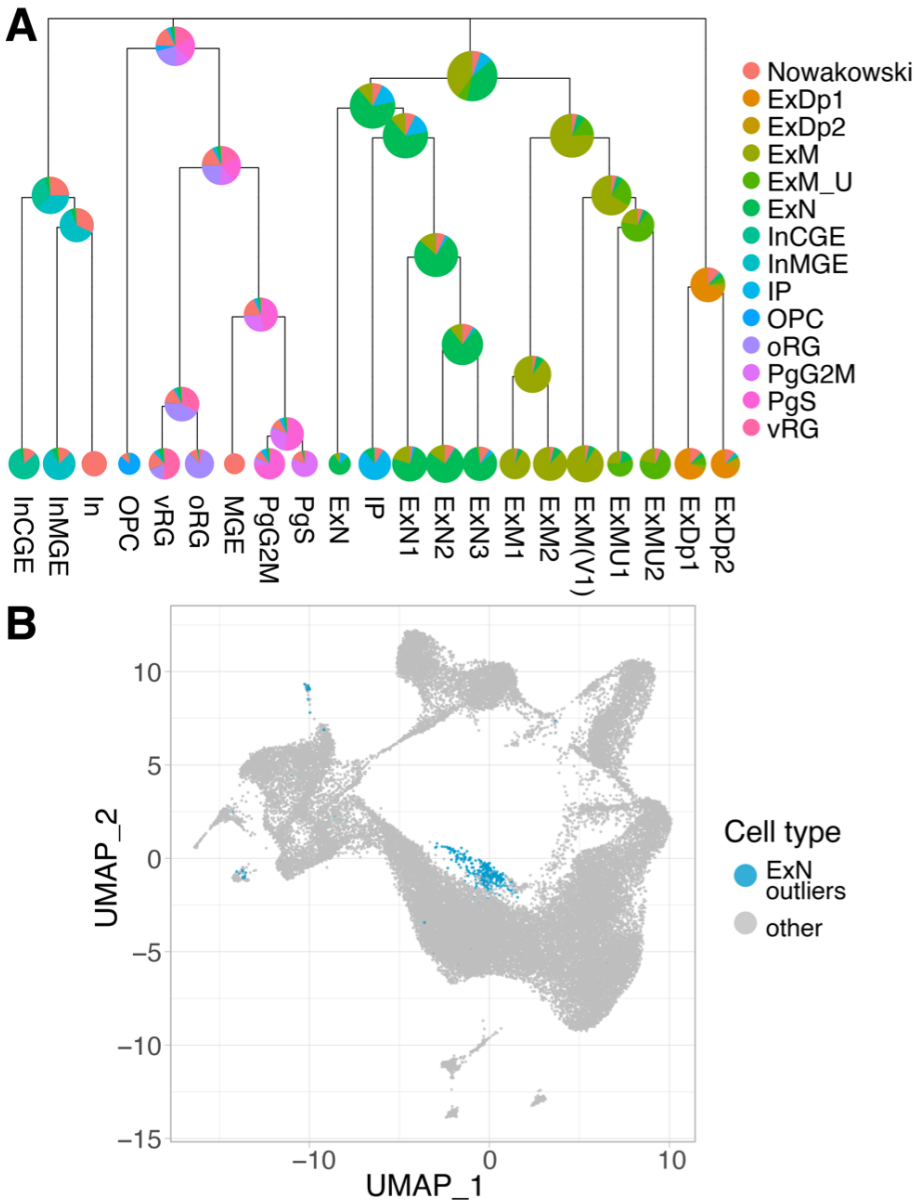
Supplementary Fig. 6. *By using the provided per-gene p-values supplied by Kaplanis et al. (denovoWEST_p_full), we can transform the p-values into a FDR (FDR denovoWEST). Comparing the FDRs from TADA-DD and denovoWEST yields strong concordance between significance measures from the two approaches, attaining a correlation coefficient of 0.95 on the log10 scale. The points in red are genes for which the TADA-DD BF evidence of association is so large that they exceed computational precision, and their FDR computationally returned 0. For these genes, the smallest non-zero TADA-DD FDR was substituted as a proxy for this figure.*



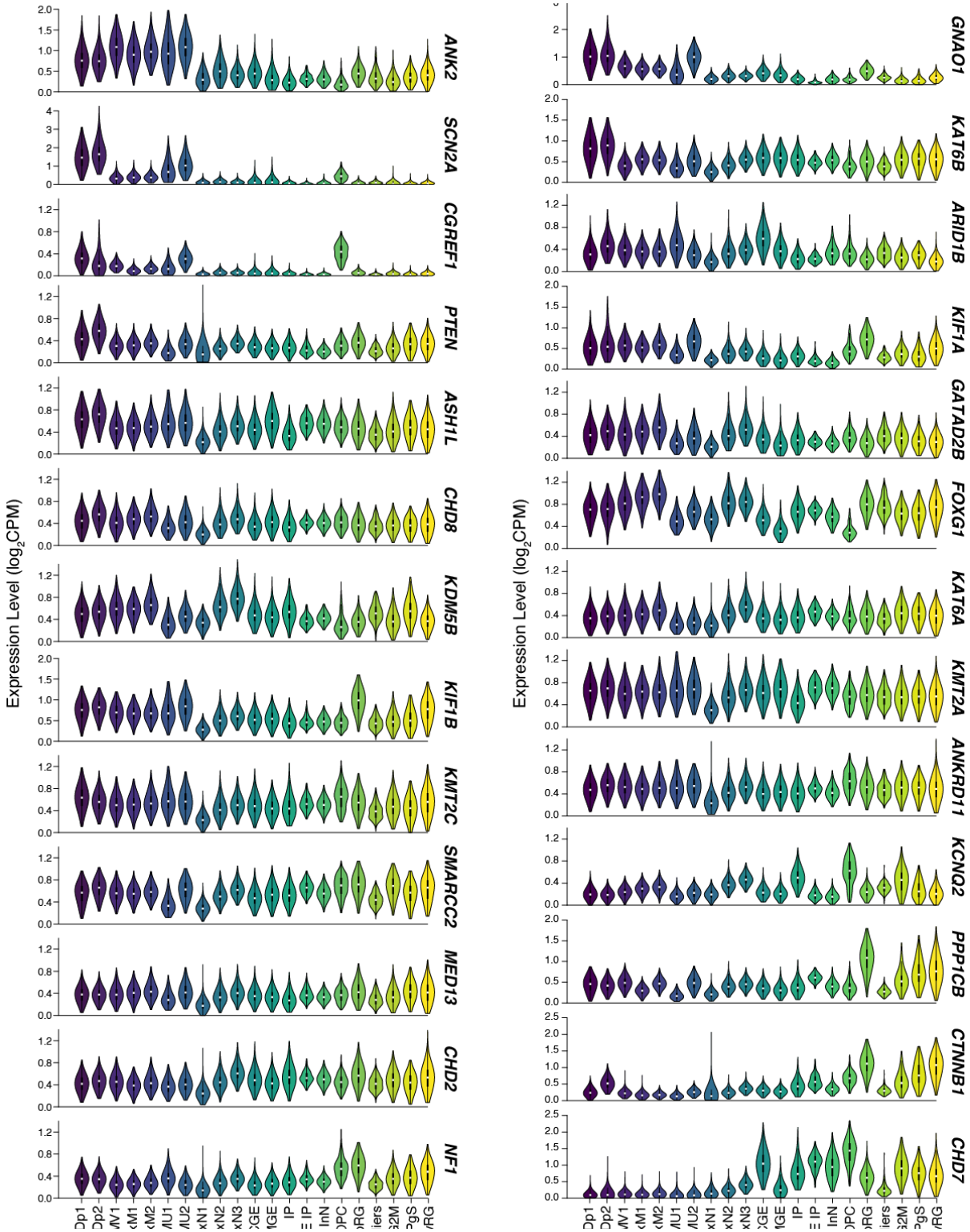
Supplementary Fig. 7. *For the set of 373 TADA-NDD genes with FDR ≤ 0.001, we examined the concordance of BF evidence between cohorts and cohort subcomponents. The ASD cohort is composed of two halves of roughly equal sample sizes - the ASC+SSC portion and the SPARK portion. We considered only evidence from de novo PTVs, MisB, and MisA variants for calculating BF evidence so that the modes of evidence considered could be identical between the ASD cohort and the DD cohort. The DD cohort was down-sampled to match the total ASD cohort in size. Pairwise comparison of log10 BFs between ASC+SSC and SPARK was plotted for the 373 genes and achieved a correlation of 0.78, while pairwise comparison of log10 BFs between the combined ASD cohort and the DD cohort achieved a correlation of 0.42.*



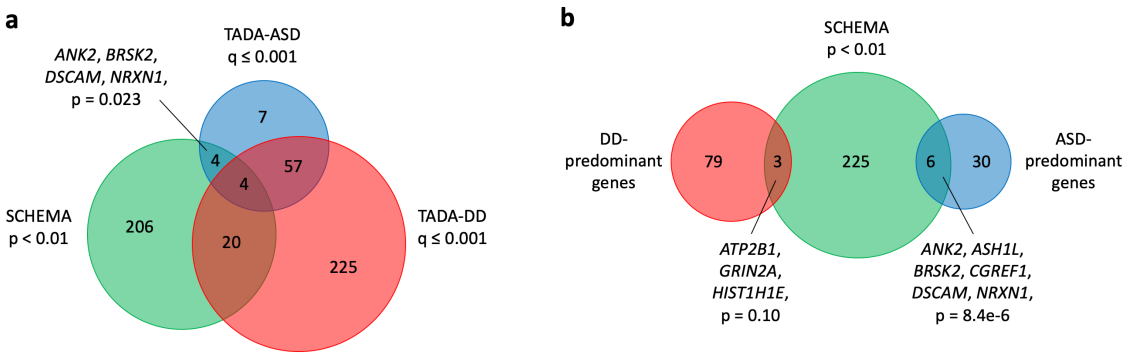
Supplementary Fig. 8. *UMAP of scaled factor loadings obtained from data integration. On the left, only cells from the Polioudakis data are colored into 16 major cell types. Similarly, on the right, only cells from Nowakowski data are colored according to the 48-cell-type label. See SI Appendix, Table S2*[74] *for detailed cell-type annotations from respective studies.*



Supplementary Fig. 9. a, *Results of MRtree applied to the integrated data, resulting in 21 clusters. Nodal labels are spelled out in Fig. 5. Each node is a pie chart showing the fraction of cells from the Nowakowski data.* ***b,*** *The cluster indicated in blue was considered an outlier and removed from downstream analysis. This small, outlying cluster, originally labeled as ExN by Polioudakis, did not cluster with the other ExN cell types.*



Supplementary Fig. 10. *Expression data from the single-cell dataset for ASD-predominant genes (left) and DD-predominant genes (right) selected as the extremes of the posterior probability metric (**Supplemental Table 15** and **Fig. 5f**). Genes are ordered by visual appearance from neuron-enriched (top) to progenitor-enriched (bottom).*



Supplementary Fig. 11. Overlap of schizophrenia-associated genes with ASD- and DD-associated genes. a, *Venn diagram showing overlap among ASD-associated genes, DD-associated genes, and schizophrenia-associated genes.* ***b,*** *Venn diagram showing overlap among ASD-predominant genes, DD-predominant genes, and schizophrenia-associated genes.*

Supplemental information on Statistical Tests

Fig. 1b,c,e,f: Due to the fact that many individuals harbor more than one inherited variant, we chose to calculate and display the excess of mutations in each inheritance category using a difference in rates (instead of odds ratios) so that all measures are on a comparable scale across *de novo*, case/control, and inherited. For *de novo* variants, the binomial test statistic is the ratio of the *de novo* variants in the affected offspring divided by the rate of *de novo* variants in all offspring, with the null probability equal to the number of affected offspring over all offspring. For case/control, the test statistic and null probability are constructed with the equivalent case/control numbers. For the inherited variants, the binomial test statistic is the proportion of inherited variants, and the null probability is 0.5.

Page 4
Comparing odds ratio of deletions and PTVs over first decile of LOEUF: We set the null hypothesis to be that the enrichment of deletions is the same as the enrichment in PTVs. We then randomly generate the number of deletions observed in affected offspring using the null hypothesis enrichment, multiplied by the base observed rate of mutations in unaffected offspring, to calculate an odds ratio. This process is repeated 10^7 times, and the p-value is calculated as the fraction of iterations where the odds ratio exceeded the observed odds ratio.

Page 5
Logistic regression: For each comparison noted with a logistic regression comparison, we modeled ASD status (or male versus female where appropriate) as the outcome variable and the two variables of interest as input variables. Regression outputs are then reported.

Miscellaneous
Correlation tests: R function `ggpubr::cor.test` was used to evaluate the significance of correlation coefficients where reported.

Supplementary Note References

72. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
73. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
74. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
75. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
76. Chaste, P. *et al.* A genome-wide association study of autism using the Simons Simplex Collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biol. Psychiatry* **77**, 775–784 (2015).
77. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
78. Du, H. *et al.* Transcription Factors Bcl11a and Bcl11b Are Required for the Production and Differentiation of Cortical Projection Neurons. *Cereb. Cortex* (2021) doi:10.1093/cercor/bhab437.
79. Tahirovic, S. *et al.* Rac1 regulates neuronal polarization through the WAVE complex. *J. Neurosci.* **30**, 6930–6943 (2010).
80. Katoh, H., Hiramoto, K. & Negishi, M. Activation of Rac1 by RhoG regulates cell migration. *J. Cell Sci.* **119**, 56–65 (2006).
81. Li, M., Fan, Y., Wang, Y., Xu, J. & Xu, H. ZMIZ1 promotes the proliferation and migration of melanocytes in vitiligo. *Exp. Ther. Med.* **20**, 1371–1378 (2020).
82. Carapito, R. *et al.* ZMIZ1 Variants Cause a Syndromic Neurodevelopmental Disorder. *Am. J. Hum. Genet.* **104**, 319–330 (2019).
83. Heng, Y. H. E., Barry, G., Richards, L. J. & Piper, M. Nuclear factor I genes regulate neuronal migration. *Neurosignals* **20**, 159–167 (2012).
84. Wong, Y. W. *et al.* Gene expression analysis of nuclear factor I-A deficient mice indicates delayed brain maturation. *Genome Biol.* **8**, R72 (2007).
85. Hou, P.-S., hAilín, D. Ó., Vogel, T. & Hanashima, C. Transcription and Beyond: Delineating FOXG1 Function in Cortical Development and Disorders. *Front. Cell. Neurosci.* **14**, 35 (2020).
86. Birkhoff, J. C., Huylebroeck, D. & Conidi, A. ZEB2, the Mowat-Wilson Syndrome Transcription Factor: Confirmations, Novel Functions, and Continuing Surprises. *Genes* **12**, (2021).
87. Jung, E.-M. *et al.* Arid1b haploinsufficiency disrupts cortical interneuron development and mouse behavior. *Nat. Neurosci.* **20**, 1694–1707 (2017).
88. McKinsey, G. L. *et al.* Dlx1&2-dependent expression of Zfhx1b (Sip1, Zeb2) regulates the fate switch between cortical and striatal interneurons. *Neuron* **77**, 83–98 (2013).
89. Mutch, C. A., Schulte, J. D., Olson, E. & Chenn, A. Beta-catenin signaling negatively regulates intermediate progenitor population numbers in the developing cortex. *PLoS One* **5**, e12376 (2010).
90. Gao, Y. *et al.* Loss of histone methyltransferase ASH1L in the developing mouse brain causes autistic-like behaviors. *Commun Biol* **4**, 756 (2021).
91. Jabaudon, D., Shnider, S. J., Tischfield, D. J., Galazo, M. J. & Macklis, J. D. RORβ induces barrel-like neuronal clusters in the developing neocortex. *Cereb. Cortex* **22**, 996–1006 (2012).
92. Yang, R. *et al.* ANK2 autism mutation targeting giant ankyrin-B promotes axon branching and ectopic connectivity. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15262–15271 (2019).
93. van Woerden, G. M. *et al.* TAOK1 is associated with neurodevelopmental disorder and essential for neuronal maturation and cortical development. *Hum. Mutat.* **42**, 445–459

(2021).
94. Spratt, P. W. E. *et al.* The Autism-Associated Gene Scn2a Contributes to Dendritic Excitability and Synaptic Function in the Prefrontal Cortex. *Neuron* **103**, 673–685.e5 (2019).
95. Südhof, T. C. Synaptic Neurexin Complexes: A Molecular Code for the Logic of Neural Circuits. *Cell* **171**, 745–769 (2017).
96. Inagaki, N. *et al.* CRMP-2 induces axons in cultured hippocampal neurons. *Nat. Neurosci.* **4**, 781–782 (2001).