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

Jack Fu, F. Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan Collins, Shan Dong, Brie Wamsley, Lambertus Klei, Lily Wang, Stephanie Hao, et al.

▶ **To cite this version:**

Jack Fu, F. Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan Collins, et al.. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nature Genetics, 2022, 54 (9), pp.1320-1331. 10.1038/s41588-022-01104-0 . inserm-03949950

HAL Id: inserm-03949950

https://inserm.hal.science/inserm-03949950

Submitted on 20 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Rare coding variation illuminates the allelic architecture, cellular expression patterns, and phenotypic context of autism

Jack M. Fu[1,2,3,*], F. Kyle Satterstrom[2,4,5,*], Minshi Peng[6,*], Harrison Brand[1,2,3,7,*], Ryan L. Collins[1,2,3,8], Shan Dong[9], Brie Wamsley[10], Lambertus Klei[11], Lily Wang[2,8], Stephanie P. Hao[1,3,7], Christine R. Stevens[2,4,5], Caroline Cusick[4], Mehrtash Babadi[12], Eric Banks[12], Brett Collins[13,14,15], Sheila Dodge[16], Stacey B. Gabriel[16], Laura Gauthier[12], Samuel K. Lee[12], Lindsay Liang[9], Alicia Ljungdahl[9], Behrang Mahjani[13,14,17], Laura Sloofman[13,14,15], Andrey N. Smirnov[12], Mafalda Barbosa[15,18], Catalina Betancur[19], Alfredo Brusco[20,21], Brian H.Y. Chung[22], Edwin H. Cook[23], Michael L. Cuccaro[24], Enrico Domenici[25], Giovanni Battista Ferrero[26], J. Jay Gargus[27], Gail E. Herman[28], Irva Hertz-Picciotto[29], Patricia Maciel[30], Dara S. Manoach[31], Maria Rita Passos-Bueno[32], Antonio M. Persico[33], Alessandra Renieri[34,35,36], James S. Sutcliffe[37,38], Flora Tassone[29,39], Elisabetta Trabetti[40], Gabriele Campos[32], Simona Cardaropoli[26], Diana Carli[26], Marcus C.Y. Chan[22], Chiara Fallerini[34,35], Elisa Giorgio[20], Ana Cristina Girardi[32], Emily Hansen-Kiss[41], So Lun Lee[22], Carla Lintas[42], Yunin Ludena[29], Rachel Nguyen[27], Lisa Pavinato[20], Margaret Pericak-Vance[24], Isaac N. Pessah[29,43], Rebecca J. Schmidt[29], Moyra Smith[27], Claudia I.S. Costa[32], Slavica Trajkova[20], Jaqueline Y.T. Wang[32], Mullin H.C. Yu[22], The Autism Sequencing Consortium (ASC), Broad Institute Center for Common Disease Genomics (Broad-CCDG), iPSYCH-BROAD Consortium, David J. Cutler[44], Silvia De Rubeis[13,14,15,45], Joseph D. Buxbaum[13,15,46,+], Mark J. Daly[1,2,4,5,47,48,+], Bernie Devlin[11,+], Kathryn Roeder[6,49,+], Stephan J. Sanders[9,+], 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

*** first author**

+ corresponding author:

J.D.B - joseph.buxbaum@mssm.edu; M.J.D - mjdaly@atgu.mgh.harvard.edu; B.D. - devlinbj@upmc.edu; K.R. - roeder@andrew.cmu.edu; S.J.S. - Stephan.Sanders@ucsf.edu; M.E.T. - MTALKOWSKI@mgh.harvard.edu;

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. Interdepartmental Program "Autism 0-90", "Gaetano Martino" University Hospital, University of Messina, Messina 98125, 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

Abstract

Some individuals with autism spectrum disorder (ASD) carry functional mutations rarely observed in the general population. We explored the genes disrupted by these variants from joint analysis of protein-truncating (PTV), missense, and copy number variants (CNVs) in a cohort of 63,237 individuals. We discovered 72 ASD risk genes at false discovery rate (FDR)≤0.001 (185 at FDR≤0.05). *De novo* PTVs, damaging missense variants, and CNVs represented 57.5%, 21.1%, and 8.44% of association evidence, while CNVs conferred greatest relative risk. Meta-analysis with cohorts ascertained for developmental delay (DD, N=91,605) yielded 373 ASD/DD risk genes at FDR≤0.001 (664 at FDR≤0.05), some of which differed in relative frequency of mutation between ASD and DD. The DD-associated genes were enriched in transcriptomes of progenitor and immature neuronal cells whereas genes displaying stronger evidence in ASD were more enriched in maturing neurons and overlapped with schizophrenia-associated genes, emphasizing that these neuropsychiatric disorders share common pathways to risk.

Introduction

Autism spectrum disorder (ASD) affects approximately 2.3% of children in the United States[1]. ASD is highly heritable[2], with the majority of risk stemming from common genetic variants, each of small effect, acting additively across the genome[3]. However, in at least 10% of ASD cases, rare and *de novo* variants confer substantial risk, and exome sequencing has enabled rare coding variant studies across ASD and many related developmental and neuropsychiatric disorders[4–7]. These studies have largely focused on single nucleotide variants (SNVs) and

insertions/deletions (indels) that arise *de novo*, though modest over-transmission to ASD probands has been observed for some classes of rare variants[4,8,9]. The relative contribution of *de novo* protein truncating variants (PTVs) to risk varies significantly by ascertainment strategy: burden is greatest in individuals with developmental delay (DD), intellectual disability (ID), or multi-system congenital anomalies; moderate in individuals with ASD or isolated developmental anomalies; and lowest in schizophrenia and other neuropsychiatric disorders[4,5,7,10,11]. Hundreds of risk genes have been discovered across these disorders, with associations largely driven by phenotypic severity and cohort size[12,13].

Early microarray studies established that individuals with ASD also harbor an excess of very large copy number variants (CNVs)[14–19]. These studies identified many recurrent genomic disorder (GD) CNVs associated with syndromic features that arose due to mispairing of long homologous segments, a mechanism known as non-allelic homologous recombination (NAHR)[14,16,20,21]. Due to their high mutation rate, GDs are among the best characterized genetic risk factors across all neurodevelopmental disorders (NDDs)[6,16,21,22]. Beyond these large segments, defining the contribution of small CNVs localized to individual genes in ASD across large cohorts has been a technical challenge. With advancing technologies, structural variant (SV) discovery is now tractable from whole-genome sequencing (WGS) and has been applied to population resources[23–25], but only to relatively small ASD cohorts[26–30]. These studies, as well as long-read WGS on a small number of individuals[31,32], have revealed the mutational diversity of SVs that exist in all genomes, >99% of which were not detectable by microarray studies[32,33]. We demonstrate here that refined models of exome-based CNV discovery can capture small, rare coding CNVs with sensitivity and specificity that is comparable to indel discovery and amenable to large-scale association studies. We reasoned that joint analyses of rare coding SNVs, indels, and CNVs at the resolution of individual genes and exons in large cohorts would provide a more complete picture of allelic diversity and mutational mechanisms that impact specific genes contributing to ASD.

Discovery of risk genes can also be enhanced through the integration of functional effects of rare variation and metrics to quantify negative selection[4]. One such measure is the 'loss-of-function observed/expected upper bound fraction' (LOEUF) score[34], which is a continuous measure of selective pressure against PTVs in each gene. Similarly, the 'missense badness, PolyPhen-2, and constraint' (MPC) score[35] is a measure of the estimated deleteriousness of missense variation. In this study, we use a Bayesian statistical framework, the Transmission and *De Novo* Association (TADA) model[36], to incorporate these functional annotations into joint analyses of coding SNVs, indels, and CNVs across the largest exome-sequenced ASD and DD cohorts to date, comprising 63,237 individuals from ASD cohorts (20,627 ASD-affected individuals) and 91,605 samples from DD cohorts (31,058 DD-affected individuals). We identify hundreds of genes associated with these disorders and reveal significant overlap, as well as substantial heterogeneity, in the genes associated with each phenotype and in the neural cell types expressing them. Overall, these analyses provide new insights into the contributions of rare coding variation in NDDs, including broad overlap and nuanced distinctions of genetic risk and its influence on specific pathways and developmental trajectories.

Results

Patterns of rare coding variants in ASD

We aggregated exome sequencing data across 33 ASD cohorts that included 63,237 individuals: 15,036 affected probands, 28,522 parents, and 5,492 unaffected siblings from family data, as well as 5,591 affected and 8,597 unaffected individuals from case-control studies (**Fig.**

1a, **Supplementary Tables 1-4**). Of the family data, 58.7% had not previously been published. After filtering, variant counts were comparable across cohorts, with an average of 1.64 (1.66/affected, 1.57/unaffected) *de novo* SNVs and 0.18 (0.18/affected, 0.16/unaffected) *de novo* indels per individual. Consistent with prior studies, PTVs and damaging missense variants were enriched in individuals with ASD compared to unaffected individuals (**Fig. 1b-c**). PTV enrichment was greatest in genes under selective constraint, represented by low LOEUF scores[34] (**Supplementary Tables 5-8)**, with both *de novo* and inherited PTVs enriched in the lowest three deciles of LOEUF (binomial test, **Fig. 1b**). We annotated two groups of deleterious missense variants, MisB (MPC ≥ 2) and MisA (2 > MPC ≥ 1); MisB variants were strongly enriched in ASD cases while the effect of MisA variants was modest (**Fig. 1c**). Overall, we observed the greatest ASD risk in *de novo* variation, with less significant risk observed in rare case/control (for which *de novo* status cannot be determined) and inherited variants.

Discovery of rare and *de novo* CNVs from exome sequencing

Microarray-based studies have established a clear etiological role for large, rare CNVs in ASD[14,16–18,37–40]. Here, we applied a CNV discovery tool, GATK-gCNV, that predicts read-depth changes from short-read sequencing[41]. We performed extensive benchmarking using orthogonal technologies across 7,035 individuals with matching CNVs detected from WGS[26,42]. These analyses observed 86% sensitivity and a positive predictive value (PPV) of 90% to detect rare (site-frequency <1%) CNVs discoverable by WGS at a resolution greater than two captured exons (**Fig. 1d**), and comparable sensitivity (83%) and PPV (97%) for *de novo* CNVs (**Supplementary Figs. 1-2**). Using these site-frequency and resolution filters, we analyzed CNVs in 55,678 samples with accessible data (**Methods, Supplementary Table 4**). We observed 17,774 rare inherited and 662 *de novo* autosomal CNVs after filtering; 3.95% of ASD cases and 1.39% of unaffected siblings harbored at least one *de novo* coding CNV (odds ratio [OR]: 2.91, $p=2.2\times10^{-21}$, Fisher's exact test; **Fig. 1e-f**, **Supplementary Table 9**). A greater proportion of female cases harbored *de novo* CNVs than males (6.0% vs. 3.5%, OR: 1.8, $p=2.1\times10^{-8}$, Fisher's exact test), consistent with a female protective effect that proposes a higher burden of risk factors required for an ASD diagnosis in females[17,43]. *De novo* deletions spanning at least one constrained gene (LOEUF<0.4) showed the greatest enrichment in ASD cases across all variant classes (9.33 fold enrichment, $p=6.7\times10^{-21}$, binomial test), with a relative difference approximately three-fold higher than *de novo* PTVs in the same constraint decile ($p=2.3\times10^{-4}$, permutation test). Duplications displayed similar but more attenuated enrichment patterns (**Fig. 1e-f)**.

We next sought to dissect the relative impact of large GD segment CNVs (**Fig. 2a**) from alterations to individual genes. We considered 79 GD segments previously associated with NDDs, as described in Collins *et al.*[22] (**Supplementary Table 10**). Of the 662 *de novo* CNVs discovered, 253 (38.2%) matched one of these loci (**Methods**). As expected, *de novo* GDs were strongly enriched in ASD cases (deletion OR: 4.8, $p=2.6\times10^{-8}$, duplication OR: 2.9, $p=3.6\times10^{-5}$, Fisher's exact test, **Fig. 2b**), whereas a weak trend was detected for inherited GDs (OR:1.2, p=0.053, Fisher's exact test). After excluding GD segments, the remaining 409 *de novo* CNVs were enriched in ASD probands, but with more modest effect sizes (non-GD deletion OR: 3.1, $p=1.1\times10^{-9}$; non-GD duplication OR: 2.1, $p=5.4\times10^{-4}$, Fisher's exact test). However, the impact of a non-GD *de novo* deletion of a constrained gene was comparable to a GD deletion (OR: 6.9, $p=2.2\times10^{-12}$, Fisher's exact test, **Fig. 2c**) and significantly greater than *de novo* PTVs in constrained genes (OR: 2.74, $p=3.7\times10^{-34}$, Fisher's exact test, **Fig. 2c**).

We also quantified risk associated with GDs in ASD compared to the general population by applying GATK-gCNV to exome data in the UK Biobank (UKBB)[44]. We processed the UKBB data using identical parameters as the ASD cohort and compared carrier rates for 79 GD loci in

13,786 ASD cases and 145,532 UKBB controls with accessible phenotype information and no documented neuropsychiatric or developmental phenotypes. These analyses demonstrated a linear inverse correlation of decreasing OR in ASD with increasing GD frequency in the UKBB, with the most significant loci including established GDs such as 15q11.2-q13.1, 17q12, and 17q11.2, among others (OR>50; **Fig. 2d**). We provide these results in **Supplementary Fig. 3** and **Supplementary Table 10** as a reference for future GD variant interpretation.

Finally, *De novo* SNVs and indels more frequently arise on the paternal allele[38,45,46], yet a maternal bias has been observed for *de novo* CNVs in ASD[47]. We explored the mechanisms associated with this bias using SNV/indel data to estimate the parent of origin for 225 *de novo* CNVs and observed no bias in ASD cases (**Fig. 2e**, 49% maternal, p=0.89, binomial test; **Methods**). However, 69% of *de novo* CNVs at NAHR-mediated GD preferentially arose on the maternal allele ($p = 3.7x10^{-4}$, binomial test) and recapitulated prior findings, with the strongest bias observed for the 16p11.2 CNV across this cohort and the Simons Searchlight project[47,48] (95% maternal origin, **Fig. 2e**). By contrast, CNVs that were not NAHR-mediated GDs showed a significant paternal bias (63.5%; $p = 2.0x10^{-3}$, binomial test), suggesting a mechanistic maternal bias in NAHR-mediated CNV formation, but a paternal bias in all other classes of *de novo* structural variants, consistent with prior analyses using WGS[42].

Integration of variant classes for ASD gene discovery

The relative risk of variants associated with ASD varied by mode of inheritance, variant class (PTV, MisB, MisA, deletion, duplication), and evolutionary constraint. We thus sought to leverage these insights to refine ASD gene discovery by extending a Bayesian analytic framework, TADA[4,36], to include: (1) rare and *de novo* CNVs, (2) variants present in unaffected offspring, and (3) evolutionary constraint from gnomAD (LOEUF[34]**, Methods, Supplementary Table 8, Supplementary Fig. 4**). For each autosomal protein-coding gene, a Bayes Factor (BF) was calculated to represent evidence of association across variant types and modes of inheritance, taking into account mutation rates and prior relative risks (**Fig. 3a**).

Applying this model to the aggregated ASD data (TADA-ASD), we identified 72 genes associated with ASD at FDR ≤ 0.001 (**Fig. 3b**) and 185 genes at FDR ≤ 0.05 (**Supplementary Table 11**). Within the 72 genes, *de novo* PTV, MisB, or MisA variants were detected in 4.0% of cases and 0.5% of controls (combined OR: 8.44, $p = 3.4x10^{-51}$, Fisher's exact test), and we applied cross-validation to refine variant class-specific risk (**Supplementary Note**, **Supplementary Table 12**). Notably, the FDR ≤ 0.001 used here is approximately equivalent to an exome-wide Bonferroni significance threshold ($p < 2.8x10^{-6}$) when back-calculating a p-value and correcting for 18,128 autosomal genes, making it comparable to recent studies of schizophrenia[7] and DD[5]. We calibrated the relative impact of the inclusion of multiple variant classes and our updated model parameters here compared to prior ASD studies on a subset of these samples (**Fig. 3c**, **Supplementary Fig. 5**). While we observed considerable mutational diversity across ASD risk genes (**Fig. 3d-e**), haploinsufficiency was the predominant mechanism; PTVs and deletions accounted for >90% of the evidence in 21 of 72 ASD risk genes (29.2%). However, for 9 genes (12.5%), >90% of evidence was derived from missense variants and duplications (e.g., *DEAF1, SLC6A1*; **Fig. 4a**), including one gene (*PLXNA1*) where over-transmission of missense variants was observed specifically within the Plexin domain of the encoded protein (**Fig. 4b-c**).

Although this framework is not intended to assess autosomal recessive risk in ASD, we examined offspring with two (or more) PTV and/or MisB alleles within the same gene, whether from homozygous or compound heterozygous variants. We found 10 genes with two or more occurrences in ASD cases (*B3GALT6*, *BTN2A2*, *DNAAF3*, *EIF3I*, *FEV*, *KCP*, *RDH11*, *RNF39*,

RNF175, and *SSPO*), and no such occurrence in unaffected siblings. Some genes such as *FEV* have been implicated in recessive models of ASD[49], whereas *de novo* missense variants in *EIF3I* have been associated with a neurodevelopmental syndrome but not yet an autosomal recessive form of ASD[50].

Lastly, we evaluated two hypotheses regarding the excess burden of *de novo* variants in females across the 185 FDR ≤ 0.05 genes and the GD loci: (1) the excess is due to a female protective effect; or (2) it arises from an ascertainment bias by which females diagnosed with ASD tend to be more severely affected than males[51,52]. In fact, both severity and sex are associated with being a carrier of such mutations. Using dichotomized full-scale IQ (FSIQ, 2,095 samples) and autism diagnostic observational scale (ADOS, 5,280 samples) test scores as proxies of phenotypic severity, we constructed logistic regressions to estimate the odds ratio of carrying a *de novo* damaging variant (PTV, MisB, GD CNV, or CNV overlapping one of the 185 ASD genes) as a function of sex and phenotype. We found that ASD individuals harboring these damaging mutations are significantly more likely to be female and to be severely affected, and that sex and severity status combined additively to determine burden. There was no evidence of an interaction effect, which would be expected with ascertainment bias (**Methods, Supplementary Table 13a-d**). Thus, these analyses strongly favor the female protective effect.

Comparing the genetic architectures of ASD and general DD

Significant overlap has been observed between genes affecting ASD and those affecting development more broadly, including NDDs[53,54]. To explore commonalities and differences across genes that impact NDD risk, we sought to integrate data from our ASD cohort with an independent cohort of 31,058 offspring ascertained for broadly defined DD and their parents[5]. *De novo* SNVs and indels from this cohort were recently analyzed using DeNovoWEST, a permutation-based frequentist method, which reported association for 252 autosomal genes[5]. We re-analyzed these data using our TADA framework to enable direct comparisons between cohorts using uniform statistical models and significance thresholds. This implementation identified 309 autosomal genes associated at FDR ≤ 0.001 (TADA-DD), including 237 (94%) of the 252 autosomal genes discovered previously[5] (**Supplementary Table 11**). Moreover, our FDR values were highly correlated with those derived from the DeNovoWEST significance values[5] ($r=0.95$, $p<1.0\times10^{-22}$, **Supplementary Fig. 6**). As expected, given the enrichment of cases with severe and syndromic disorders in the DD cohort[5], the *de novo* PTV, MisB, and MisA counts in offspring showed similar but much stronger variant enrichment across the top three deciles of LOEUF (**Fig. 5a-b**).

Because a cardinal rule of meta-analysis is that the data should not be too heterogeneous, before combining results across cohorts, we assessed whether the genes identified in the ASD cohort were also associated in the DD cohort, and vice versa. To do so, we converted the distribution of TADA FDRs to p-values for each study (**Methods**). If the genes associated in one cohort were also associated in the other, or some fraction thereof, the distribution of their association p-values would be skewed toward zero. When we selected the 477 genes associated in the DD cohort from the TADA-DD analysis at FDR ≤ 0.05, the estimated fraction of ASD genes also showing association was 0.701 (**Methods**, **Fig. 5c**), indicating that 70.1% of these DD genes affect risk for ASD. The converse conditioning estimated that 86.6% of ASD risk genes have broad effects on development (**Fig. 5c-d**). Thus, because the ASD and DD cohorts are somewhat complementary, we conducted a joint analysis using the TADA framework to integrate the genetic evidence for each gene across the cohorts by combining the BFs, conceptually similar to a frequentist meta-analysis. This combined analysis (TADA-NDD) revealed 373 genes associated with general NDDs at FDR ≤ 0.001 (664 genes at FDR ≤ 0.05; **Supplementary Table 11**). Notably, 54 of the 373 genes did not achieve FDR ≤ 0.001 in either

cohort alone, demonstrating a 14% increase in yield. Although we did not have access to CNV data from the DD cohort, we nonetheless found a profound and specific enrichment of 134 *de novo* CNVs that impacted one of the 373 TADA-NDD genes across all ASD cases and only one such CNV in siblings (OR: 48.9, $p=6.4x10^{-17}$, Fisher's exact test). We also used this set of genes to assess support for an oligogenic model of ASD and DD, finding no support for the hypothesis (**Methods**, **Supplementary Tables 14a-c**).

Heterogeneity of mutation patterns between ASD/DD risk genes

Isolating genes that exert a greater effect on ASD than they do on other DDs has remained challenging due to the frequent comorbidity of these phenotypes. Still, an estimated 13.4% of the TADA-ASD genes show little evidence for association in the DD cohort (**Fig. 5d**). The remainder are likely pleiotropic, yet some could have a greater impact on ASD risk than other features of development. To evaluate heterogeneity between the ASD and DD cohorts, we retained only *de novo* SNVs/indels for independent gene-level BF calculations. For the 373 genes at TADA-NDD FDR ≤ 0.001, we observed a Pearson's correlation of 0.78 of the gene-level log BF between the two major ASD sub-cohorts (the Simons Powering Autism Research [SPARK] initiative versus all others) compared to 0.42 between the ASD and DD cohorts, reflecting more consistent evidence between ASD cohorts than between ASD and DD cohorts (**Supplementary Fig. 7**).

We next determined which genes were more commonly mutated in one cohort or the other by selecting 464 "signal genes" (**Supplementary Table 15**). These genes were defined as any gene with FDR ≤ 0.05 in either TADA-ASD or TADA-DD from *de novo* PTVs and MisB variants, which as classes confer similar relative risk for ASD (**Fig. 1b, c**); MisA variants were excluded because they conferred far less risk (**Fig. 1b, c**). Of these signal genes, 120 belonged to TADA-ASD, 428 to TADA-DD, and 84 to both. Notably, the 84 genes significant in both cohorts still demonstrated significant variant count heterogeneity ($X^2=317.6$, DF=83, $p=3.8x10^{-23}$) between the cohorts. A common way to assess which of the 464 genes have more variation in either cohort would be a standardized chi-squared test statistic (C statistic, **Methods**), but its power to discriminate is abrogated by the much higher burden of risk variants in the DD cohort (**Fig. 5a-b**). We therefore adjusted for the difference in mutational burden between the cohorts by randomly downsampling the DD mutations to be comparable to that for ASD mutations. A mixture model was then adopted to disentangle the two commingled distributions, assigning posterior probabilities that a gene is from the ASD or DD component of the statistical distribution (**Fig. 5e-f, Supplementary Table 15**). Using a posterior probability cutoff of greater than 0.99, we find 36 genes to be a part of the ASD mixture component (ASD-predominant) and 82 genes to be a part of the DD component (DD-predominant) (**Fig. 5f, Supplementary Table 15**).

Differential neuronal layers impacted by ASD/DD risk genes

To explore differences in expression between genes identified across ASD and DD cohorts, we examined single-cell gene expression patterns from human fetal brains. Two studies provided data from more than 37,000 cortical cells ranging from 6-27 weeks post-conception[55] (**Supplementary Table 16**). To combine these datasets, we adjusted for batch effects using cFIT[56]. UMAP plots showed that similar cell types from the different batches grouped together, while cells unique to either batch were preserved (**Fig. 6a**, **Supplementary Fig. 8**). We applied unsupervised clustering to the combined data to identify cell subtypes in the context of a hierarchical tree to illustrate the relationships between major and minor cell type clusters. Using the MRtree method[57], we observed that cells of each labeled type were merged across datasets into common clusters. Visualizing the tree, the major branches corresponded to glial and progenitor cells, excitatory neurons, deep layer enriched excitatory neurons, and inhibitory neurons (**Supplementary Table 17**). Likewise, minor splits reflected expected relationships

between cell types (**Supplementary Fig. 9a**). Based on the trajectory analysis of Polioudakis e*t al*.[58], the ExN clade is less differentiated than the ExM clade, which in turn is less differentiated than the ExMU clade.

Next, we assessed the enrichment of ASD and DD risk genes meeting the posterior probability 0.99 threshold within cell clusters (**Fig. 5f**). Among the 36 genes classified as ASD-predominant, 22 were expressed in these cell types; of the 82 genes classified as DD-predominant, 59 were expressed. Using odds ratios to reflect the strength of signal, both ASD-predominant and DD-predominant genes were enriched in interneurons and excitatory neurons compared to glial and progenitor cells (**Fig. 6b, Supplementary Fig. 10, Supplementary Tables 15, 18-19**). ASD-predominant enrichment appeared somewhat stronger than DD-predominant enrichment in excitatory neuron lineages, with a difference in log odds (comparing enrichment in the major clade of excitatory neurons to progenitors) of 1.29 for ASD and 0.7 for DD (one-sided p = 0.017 for ASD; p = 0.031 for DD).

The DD-predominant expressed genes tend to occur in cell types that are less differentiated than the corresponding cell type enriched for ASD-predominant genes: ExN3, ExM2, IP, InCGE (for details see **Supplementary Table 18** and **Supplementary Note**). By contrast, ASD-predominant expressed genes (**Supplementary Table 19**) are strongly enriched in only one cell type, maturing excitatory neurons (ExMU1) and its clade. These genes highlight a shift from mainly migration-focused genes to more mature processes involved in building the neurons' nascent connectivity. If we judge enrichment solely by significance after Bonferroni correction for 21 cell types, ExMU1 remained significant for enrichment of ASD-predominant genes; likewise, ExN3 remained significant for enrichment of DD-predominant genes. Our results are consistent with DD-predominant genes being expressed earlier in development and in less differentiated cells than ASD-predominant genes.

Emergence of shared risk genes in schizophrenia and ASD

Shared genetic risk between ASD and schizophrenia, as well as other neuropsychiatric disorders, has long been postulated[59]. The Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium recently identified 244 genes associated with schizophrenia at $p < 0.01$ [7], 234 of which are in our TADA model. Among the 72 ASD genes we discovered at FDR ≤ 0.001, 61 were associated with DD (using TADA-DD FDR ≤ 0.001), and eight were associated with schizophrenia at $p < 0.01$. These two groups of 61 ASD/DD genes and eight ASD/schizophrenia genes overlap each other less than expected (p = 0.023, binomial test, **Methods**, **Supplementary Fig. 11a**). Similarly, using the gene sets shown in **Fig. 5f,** six of the 36 ASD-predominant genes (*ANK2*, *ASH1L*, *BRSK2*, *CGREF1*, *DSCAM*, and *NRXN1*) are schizophrenia-associated, while only three of the 82 DD-predominant genes (*ATP2B1*, *GRIN2A*, and *HIST1H1E*) are schizophrenia-associated. The ASD-schizophrenia overlap was significantly enriched ($p = 8.4 \times 10^{-6}$, binomial test), while the DD-schizophrenia overlap was not (p = 0.10, binomial test, **Methods**, **Supplementary Fig. 11b**). The two outcomes (6/36 vs. 3/82) were also different when compared to each other (p = 0.023, Fisher's exact test). Together, these data suggest that one subset of ASD risk genes may overlap DD while a different subset overlaps schizophrenia.

Discussion

Integrating rare protein-coding SNVs, indels, and CNVs across 63,237 individuals from ASD cohorts reveals an allelic spectrum of rare coding variation associated with ASD that is dominated by *de novo* PTVs, damaging missense variants, and deletions of constrained genes. Nonetheless, many genes were associated with multiple inheritance or variant classes and

some displayed the strongest evidence from *de novo* missense variants and duplications. While discovery is currently driven by *de novo* variants imparting loss of function, larger ASD cohorts will likely catalyze future discoveries from the subtler and more heterogeneous functional effects of missense variants and intragenic or individual exon duplications. Independently applying the same statistical model to both the DD and ASD datasets reinforces that our analytic framework and statistical thresholds are robust, as our results for DD are highly correlated with the permutation-based approach applied to those same data[5]. Integrating the two cohorts together yielded 373 genes at FDR ≤ 0.001, including 54 genes that were unique to the joint analyses and were not captured by either dataset alone, and 664 likely risk genes at FDR ≤ 0.05.

This study is also the largest exploration to date of CNVs at the resolution of individual genes and exons to ASD architecture. Benchmarking against WGS, >85% of all rare coding CNVs spanning more than two exons could be recalled by exome-based CNV discovery. We find that deletion of a highly constrained gene confers comparable risk to alteration of an established GD segment, and we observe a dramatic enrichment of CNVs among ASD probands compared to unaffected siblings across the 373 NDD risk genes identified. We also recapitulate the observation of a maternal bias in gamete-of-origin for *de novo* CNVs in ASD probands[47] but find this enrichment to be restricted to NAHR-mediated CNVs (e.g., 95% of 16p11.2 CNVs), whereas all other mechanisms were predominantly paternal in origin and consistent with prior WGS analysis in controls[42]. These results collectively emphasize the value of routine joint analysis of all classes of genomic variation in gene discovery and the potential impact of gene level CNV analyses in diagnostic testing.

We expect these findings to shed light on the neurobiological origins of ASD. However, given the substantial overlap between the genes implicated in NDDs writ large and those implicated directly in ASD, disentangling the relative impact of individual genes on neurodevelopment and phenotypic spectra is a daunting yet important challenge. Consider two of the ASD risk genes, *ARID1B* and *DSCAM*. Both are highly associated with ASD, although statistical evidence is stronger for *ARID1B*. Yet while some individuals with mutations in *ARID1B* also have comorbid ASD, it is only one of a wide range of developmental phenotypes[60]. The profound impact of *ARID1B* on development is apparent by the contrast of *de novo* mutations in the DD and ASD cohorts: 132 carriers out of 31,058 DD probands versus nine carriers out of 15,036 ASD probands, a sevenfold higher rate in DD. This raises a challenge for neurobiologists: neurodevelopmental features associated with perturbation of *ARID1B* could be relevant to DD, yet irrelevant to ASD. Because evidence for *DSCAM* comes solely from the ASD cohort, it could be a better choice for neurobiological studies of ASD. Still, as we develop here, *DSCAM* is also involved in risk for schizophrenia, and studies such as ours continue to demonstrate the pleiotropic consequences of many such genes implicated in ASD and NDD risk. To identify the key neurobiological features of ASD will likely require convergence of evidence from many ASD genes and studies. Careful selection of candidates among the genes implicated here based on their mutational and functional features could inform these future studies. We have taken a step in that direction here as genes expressed at earlier stages of cortical development, such as progenitor genes, broadly display greater DD enrichment, while those expressed later, such as maturing neurons, lean towards ASD. This is consistent with the expectation that earlier and more generalized impairment leads to severe global DD, and later, neuron-specific impairments affect more isolated developmental domains, such as social interaction and the presence of repetitive behaviors and/or interests that typify ASD.

In conclusion, our analyses of rare coding variation illuminates the allelic diversity contributing to ASD and the shared and distinct genetic architectures between ASD and related NDDs. We further highlight differential enrichment of associated genes at different neuronal timepoints. The

consortia studies aggregated here have catalyzed a rapid evolution in genetic studies in ASD, including preliminary analyses in recent preprints that have leveraged these data for insights into gene discovery in ASD and DD datasets, and into the combined impact of rare and common variant polygenic risk across males and females[8,9,61,62]. As sample sizes rapidly expand, the analytic framework presented here will continue to yield returns in both gene discovery and improved understanding of the differential risks to disorders on the neurodevelopmental and neuropsychiatric spectrum posed by variants within these genes.

Acknowledgements
We thank all of the individuals who participated in this research. We also thank all contributing investigators to the consortia datasets used here from the Autism Sequencing Consortium (ASC), the Simons Simplex Collection (SSC), Simons Powering Autism Research for Knowledge (SPARK) project, the iPSYCH project, the Deciphering Developmental Disorders (DDD) study, and Schizophrenia Exome Meta-Analysis (SCHEMA).
This work was supported by grants from the Simons Foundation for Autism Research Initiative (SSC-ASC Genomics Consortium #574598 to S.J.S., #575097 to B.D. and K.R., #573206 to M.E.T. and M.J.D., #571009 to J.D.); the SPARK project and SPARK analysis projects (#606362 and #608540 to M.E.T., M.J.D., J.D.B., B.D., K.R., S.J.S.); SFARI (#736613 and #647371 to S.J.S.), NHGRI (HG008895 to M.J.D., S.G., M.E.T.), NIMH (MH115957 to M.E.T., MH111658 and MH057881 to B.D., MH111661 and MH100233 to J.D.B., R01 MH109900 to K.R., MH111660 to M.J.D., and MH111662 and MH100027 to S.J.S..), NICHD (HD081256 to M.E.T.), AMED (JP21WM0425007 to N.O.), and the Seaver Foundation. J.M.F. was supported by an Autism Speaks Postdoctoral Fellowship and R.L.C. was supported by NSF GRFP #2017240332. E.D was supported by Fondazione Italiana Autismo (FIA-2018/53).

Autism Sequencing Consortium
Branko Aleksic 50, Mykyta Artomov 1,2,4,69, Mafalda Barbosa 15,18, Elisa Benetti 34,35, Catalina Betancur 19, Monica Biscaldi-Schafer 59, Anders D. Børglum 51,52,53,54, Harrison Brand 1,2,3,7, 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, 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, Mark J. Daly 1,2,4,5,47,48, Silvia De Rubeis 13,14,15,45, Bernie Devlin 11, 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, Jack M. Fu 1,2,3, J. Jay Gargus 27, Sherif Gerges 1,2,4,69, Elisa Giorgio 20, Ana Cristina Girardi 32, Stephen Guter 23, Emily Hansen-Kiss 41, Gail E. Herman 28, Irva Hertz-Picciotto 29, David M. Hougaard 51,60, Christina M. Hultman 17, Suma Jacob 23, Miia Kaartinen 66, Lambertus Klei 11, Alexander Kolevzon 13,14,61, Itaru Kushima 50,72, So Lun Lee 22, Terho Lehtimäki 73, Lindsay Liang 9, Carla Lintas 42, Alicia Ljungdahl 9, Caterina Lo Rizzo 35,36, Yunin Ludena 29, Patricia Maciel 30, 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, Matthew Mosconi 78, Rachel Nguyen 27, 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, Kathryn Roeder 6,49, Stephan J. Sanders 9, Sven Sandin 13,14,17, F. Kyle Satterstrom 2,4,5, Stephen W. Scherer 67,68, Sabine Schlitt 59, Rebecca J. Schmidt 29, Lauren Schmitt 23, Katja Schneider-Momm 59, Paige M. Siper 13,14,15, Laura Sloofman 13,14,15, Moyra Smith 27, Christine R. Stevens 2,4,5, Pål Suren 81, James S. Sutcliffe 37,38, John A. Sweeney 82, Michael E. Talkowski 1,2,3,4,8, Flora Tassone 29,39, Karoline Teufel 59, Elisabetta Trabetti 40, Slavica Trajkova 20, Maria del Pilar Trelles 13,14, Brie Wamsley 10, Jaqueline Y.T. Wang 32, Lauren A. Weiss 9, Mullin H.C. Yu 22, Ryan Yuen 67

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

Author Contributions Statement

Study design: M.E.T, S.J.S, K.R., B.D., M.J.D, J.D.B, S.B.G; **Sample contribution and data generation**: M.E.T., M.J.D., J.D.B., S.D.R., S.B.G., S.D., C.C., C.R.S., M.B, A.B, B.H.Y.C., M.L.C., E.D, G.B.F., J.J.G., G.E.H., I.H., P.M., D.S.M., M.R.P., A.M.P., A.R., F.T., E.T., G.C., M.C.Y.C., C.F., E.G., A.C.G., E.H., S.L.L., C.L., Y.L., R.N., L.P., M.P-V., I.N.P., E.R., R.J.S., M.S., C.I.C.S.C., S.T., J.Y.T.W., M.H.C.Y., J.S.S., E.H.C., C.B.; **Project management of coordination**: M.E.T., S.J.S., M.J.D., J.D.B., S.D.R., L.S., B.M., C.R.S., C.S., B.C.; **Methods development and analysis**: M.E.T., S.J.S., K.R., B.D., M.J.D., D.J.C., E.B., A.N.S., M.B., S.K.L., L.G., B.W., L.K., L.W., S.P.H. S.D., R.C., H.B., M.P., F.K.S., J.M.F.; **Writing**: M.E.T., S.J.S., K.R., B.D., M.J.D., J.D.B., H.B., M.P., F.K.S., J.M.F.

Competing Interests Statement

B.M.N. is a member of the scientific advisory board at Deep Genomics and consults for Biogen, Camp4 Therapeutics Corporation, Takeda Pharmaceutical, and Biogen. C.M. Freitag has been

a consultant to Desitin and Roche and receives royalties for books on ASD, ADHD, and MDD. S.J.S. has been a consultant for, and receives funding for research from, BioMarin. M.E.T. receives research funding and/or reagents from Illumina Inc., Levo Therapeutics, and Microsoft Inc. All other authors had no competing interests.

Figure legends

Fig. 1. Overview of SNV/indel and CNV rates in ASD by mode of inheritance and constraint. a, The ASD cohort consisted of 49,049 family-based samples (15,036 cases) and 14,188 case/control samples (5,591 cases). One sample was a proband in one trio and a mother in another. **b**, The relative difference in PTV frequency between cases and unaffected controls (top) and average per-sample variant count in unaffected controls (bottom) across inheritance classes (color) and LOEUF deciles (5,446 genes in top 3 deciles of LOEUF). Using a binomial test, cases were enriched for PTVs among the most constrained genes (lower LOEUF deciles), which weakened as negative selection against PTVs was relaxed (higher LOEUF deciles). **c**, Equivalent analyses were performed for missense variants annotated by MPC score and synonymous variants. Synonymous variants were not enriched in cases or controls, as evaluated via binomial tests. **d**, Benchmarking of the GATK-gCNV exome CNV discovery pipeline compared against WGS on overlapping samples achieved a sensitivity of 86% and positive predictive value of 90% for rare CNVs (<1% site frequency) that spanned more than two captured exons (red line). **e**, The relative difference in variant frequency between cases and controls for deletions. Using binomial tests, we found that the enrichment of deletions overlapping genes in the lowest LOEUF decile were stronger than PTVs in the same LOEUF deciles. **f**, Equivalent analysis for duplications demonstrated a similar pattern of enrichment compared to deletions but with more subtle relative differences.
Abbreviations: PTV: protein truncating variant; CNV: copy number variant; WGS: whole genome sequencing; WES: whole exome sequencing; MisB: missense variants with MPC score ≥2; MisA: missense variants with MPC score ≥1 and <2.
Statistical tests: b-c,e-f: two-sided binomial tests with 95% confidence interval error bars shown, p-values (not corrected for multiple tests) and sample sizes located in **Supplementary Table 22**.

Fig. 2. Contribution of CNVs to ASD by mechanism and genomic location. a, CNVs included deletions (DEL) or duplications (DUP) of genomic segments and involved a subset of recurrent sites known as genomic disorder (GD) loci. GDs mediated by non-allelic homologous recombination (NAHR) harbored recurrent breakpoints localized to flanking segmental duplications, whereas non-NAHR GDs did not. **b**, De novo CNVs were highly enriched in affected cases compared to unaffected offspring (Fisher's exact test) and the effect size was greater than that observed in de novo PTVs or de novo missense variants (logistic regression). **c**, ORs for de novo GD CNVs in probands compared to unaffected siblings, a subset of which have no observed de novo CNVs in unaffected individuals in this cohort (e.g., 16p11.2 deletions, 15q11.2-q13 duplications). **d**, Analysis of all GDs (de novo and inherited) in ASD cases compared to GDs in a population-based cohort (UK Biobank) discovered using GATK-gCNV with identical parameters, with loess-smoothed bands of the 95% confidence interval of the OR in gray. **e**, Parent-of-origin analysis of de novo CNVs using binomial tests showed maternal bias for NAHR-mediated CNVs at GD regions, which was most pronounced for the 16p11.2 GD as previously described[47].
Abbreviations: CNV: copy number variant; DEL: deletion CNV; DUP: duplication CNV; NAHR: non-allelic homologous recombination; PTV: protein truncating variant; GD: genomic disorder.
Statistical tests: b-c: Fisher's exact test with 95% confidence interval plotted as error bars, p-values (not corrected for multiple-tests) and sample sizes are located in **Supplementary Table 22**; d, Fisher's exact test of carrier status in 13,786 unique ASD cases and 143,532 unique UK biobank controls, p-values (not corrected for multiple-tests) are located in **Supplementary Table 10**; e: binomial test with 95% confidence interval plotted as error bars, sample sizes and p-values (not corrected for multiple-tests) are located in **Supplementary Table 22**.

Fig. 3. Integrating variant types and inheritance classes significantly boosts association power and reveals mutational biases within candidate genes.
a, Our new implementation of the TADA model included de novo, case/control, and rare inherited modules for each variant type: PTV, MisB, MisA, deletion, and duplication. We leveraged information from ASD probands as well as unaffected siblings in evaluating the effect of de novo variants. **b**, The evidence of ASD association contributed by each variant type for each of the 72 ASD genes with FDR ≤ 0.001. Some genes were predominantly associated with missense variants and duplications (e.g., PTEN, SLC6A1), suggesting mechanisms other than haploinsufficiency. **c**, Applying TADA to our aggregated ASD dataset yielded 72 genes at FDR ≤ 0.001, compared to 32 and 19 genes at the same threshold in previous studies on a subset of the samples (Satterstrom et al. 2020 and Sanders et al. 2015, respectively). Our expanded TADA model improved the integration of available evidence of association and increased gene discovery at equivalent statistical thresholds on the same datasets. **d-e**, We quantified the relative contribution of variant class and mode of inheritance to these 72 ASD-associated genes, demonstrating that de novo PTVs and MisB variants represented the strongest contributions to the association signals.
Abbreviations: BF: Bayes factor; PTV: protein truncating variant; MisB: missense variants with MPC score ≥2; MisA: missense variants with MPC score ≥1 and <2; Del: deletion CNV; Dup: duplication CNV; Inh: inherited; CC: case/control; DN: de novo.
Statistical tests: b, Extended TADA model.

Fig. 4. Relative contribution of evidence types in ASD risk genes. a, The relative evidence of ASD association in the extended TADA model (log_{10}BF) for the 72 ASD risk genes (FDR ≤ 0.001) shown for likely loss-of-function mechanism (PTVs and deletions) on the x-axis versus variants that may act via alternative mechanisms (missense variants and duplications) on the y-axis. **b**, Plot of the relative association evidence from de novo (y-axis) versus inherited (x-axis) variation for the 72 ASD risk genes. **c**, Evidence for ASD-association for the gene PLXNA1 was derived from de novo and inherited missense variants localized to the Plexin domain at the C-terminus of the Plexin-A1 protein.
Abbreviations: BF: Bayes factor.
Statistical tests: c: Transmission disequilibrium test.

Fig. 5. Integration of ASD and DD datasets. We performed meta-analysis of the ASD cohort with 31,058 DD trios reported in Kaplanis et al. 2020 (N = 46,094 combined NDD cases). **a**, Relative difference of de novo PTVs in ASD and DD cohorts across deciles of constraint as measured by LOEUF. **b**, Relative difference of de novo missense variants in DD and ASD cohorts across categories of MPC scores. **c,** To explore overlap in association evidence across ASD and DD risk genes, we considered the 477 TADA-DD genes with FDR ≤ 0.05. We evaluated their p-value distributions converted from TADA-ASD FDRs and observed non-uniformity suggesting that, in aggregate, 70.1% of these genes also evidence of association with ASD. **d**, In the complementary analysis of the 185 TADA-ASD genes with FDR ≤ 0.05, we looked at their p-value distributions converted from TADA-DD FDRs and again observed high non-uniformity, suggesting that in aggregate 86.6% of these genes had evidence of association with DD. **e**, Using PTV and MisB variant data, we devised a chi-squared statistic, denoted the C statistic, to measure if a gene has more observed variants in one cohort relative to the other. A mixture model was used to deconvolve the commingled distributions. **f**, We transformed the fitted mixture distribution into posterior probability for ASD enrichment. Using a cutoff of <0.01, we found 82 DD-predominant genes, while using a cutoff of >0.99 we found 36 ASD-predominant genes.
Abbreviations: PTV: protein truncating variants; BF: Bayes factor; DD: developmental delay; FDR: false discovery rate.
Statistical tests: a-b: two-sided binomial test, with 95% confidence interval error bars shown, p-values (not corrected for multiple-tests) and sample sizes located in **Supplementary Table 22**; c-d: R-3.5.3 package limma_3.38.3, e-f: mixture model.

Fig. 6. Single-cell data reveals differential neuronal layers impacted by ASD and DD genes. a, A UMAP plot visualization after integrating two studies[72,73] that provided single-cell gene expression of human fetal brains consisting of 37,000 cortical cells 6-27 weeks post-conception. Similar cell types from the two batches were grouped together while preserving cells unique to either study. See

Supplementary Table 17 for unabbreviated cell type labels and classifications to “progenitor” and “neuron” types. **b**, Both ASD- and DD-predominant genes (right and left, respectively) were found to be enriched in interneurons and excitatory neurons compared to glial cells. Compared to DD-predominant genes, ASD-predominant genes were relatively more neuron-enriched than progenitor-enriched. The developmental trajectory of excitatory neurons was approximately recapitulated in the UMAP, starting with OPC and other progenitor cells and ending with maturing upper-layer enriched and deep layer excitatory neurons. For interneurons, InCGE and InMGE were precursors to InN.
Abbreviations: UMAP: Uniform Manifold Approximation and Projection
Statistical tests: b, Fisher’s exact test

References

1. 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, 2018. *MMWR Surveill. Summ.* **70**, 1–16 (2021).
2. Sandin, S. *et al.* The Heritability of Autism Spectrum Disorder. *JAMA* **318**, 1182–1184 (2017).
3. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
4. 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).
5. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
6. Coe, B. P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
7. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* 1–9 (2022).
8. Wilfert, A. B. *et al.* Recent ultra-rare inherited mutations identify novel autism candidate risk genes. *bioRxiv* 2020.02.10.932327 (2020) doi:10.1101/2020.02.10.932327.
9. Zhou, X. *et al.* Integrating de novo and inherited variants in over 42,607 autism cases identifies mutations in new moderate risk genes. *bioRxiv* (2021) doi:10.1101/2021.10.08.21264256.
10. Lowther, C. *et al.* Systematic evaluation of genome sequencing as a first-tier diagnostic test for prenatal and pediatric disorders. *bioRxiv* 2020.08.12.248526 (2020) doi:10.1101/2020.08.12.248526.
11. Lord, J. *et al.* Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. *Lancet* **393**, 747–757 (2019).
12. Turner, T. N. & Eichler, E. E. The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. *Trends Neurosci.* **42**, 115–127 (2019).
13. Moyses-Oliveira, M., Yadav, R., Erdin, S. & Talkowski, M. E. New gene discoveries highlight functional convergence in autism and related neurodevelopmental disorders. *Curr. Opin. Genet. Dev.* **65**, 195–206 (2020).
14. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
15. Talkowski, M. E. *et al.* Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525–537 (2012).
16. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nat. Genet.* **43**, 838–846 (2011).
17. 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).
18. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
19. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94**, 677–694 (2014).
20. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
21. Lupski, J. R. Genomic disorders ten years on. *Genome Med.* **1**, 42 (2009).
22. Collins, R. L. *et al.* A cross-disorder dosage sensitivity map of the human genome. *medRxiv* (2021).

23. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* 2021.02.06.430068 (2021) doi:10.1101/2021.02.06.430068.
24. Mills, R. E. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
25. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
26. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
27. Brandler, W. M. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
28. Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
29. Turner, T. N. *et al.* Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**, 710–722.e12 (2017).
30. Ruzzo, E. K. *et al.* Inherited and De Novo Genetic Risk for Autism Impacts Shared Networks. *Cell* **178**, 850–866.e26 (2019).
31. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
32. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
33. Zhao, X. *et al.* Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am. J. Hum. Genet.* **108**, 919–928 (2021).
34. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
35. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *Cold Spring Harbor Laboratory* 148353 (2017) doi:10.1101/148353.
36. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
37. Sanders, S. J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).
38. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
39. Glessner, J. T. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459**, 569–573 (2009).
40. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* **29**, 512–520 (2011).
41. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *Cold Spring Harbor Laboratory* 201178 (2017) doi:10.1101/201178.
42. Belyeu, J. R. *et al.* De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am. J. Hum. Genet.* **108**, 597–607 (2021).
43. Robinson, E. B., Lichtenstein, P., Anckarsäter, H., Happé, F. & Ronald, A. Examining and interpreting the female protective effect against autistic behavior. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 5258–5262 (2013).
44. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
45. Dong, S. *et al.* De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9**, 16–23 (2014).
46. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
47. Duyzend, M. H. *et al.* Maternal Modifiers and Parent-of-Origin Bias of the Autism-Associated 16p11.2 CNV. *Am. J. Hum. Genet.* **98**, 45–57 (2016).
48. Simons Vip Consortium. Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders. *Neuron* **73**, 1063–1067 (2012).

49. Doan, R. N. *et al.* Recessive gene disruptions in autism spectrum disorder. *Nat. Genet.* **51**, 1092–1098 (2019).
50. Sarra, S. De novo missense variants in EIF3I cause a novel neurodevelopmental disorder with midline brain defects and skeletal abnormalities: ITHACA Call for collaborative clinical research (24). *ERN ITHACA* https://ern-ithaca.eu/de-novo-missense-variants-in-eif3i-cause-a-novel-neurodevelopmental-disorder-with-midline-brain-defects-and-skeletal-abnormalities-ithaca-call-for-collaborative-clinical-research-24/ (2020).
51. Tsirgiotis, J. M., Young, R. L. & Weber, N. A Mixed-Methods Investigation of Diagnostician Sex/Gender-Bias and Challenges in Assessing Females for Autism Spectrum Disorder. *J. Autism Dev. Disord.* (2021) doi:10.1007/s10803-021-05300-5.
52. Russell, G., Steer, C. & Golding, J. Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Soc. Psychiatry Psychiatr. Epidemiol.* **46**, 1283–1293 (2011).
53. Doshi-Velez, F., Ge, Y. & Kohane, I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics* **133**, e54–63 (2014).
54. Sanders, S. J. *et al.* A framework for the investigation of rare genetic disorders in neuropsychiatry. *Nat. Med.* **25**, 1477–1487 (2019).
55. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
56. Peng, M., Li, Y., Wamsley, B., Wei, Y. & Roeder, K. Integration and transfer learning of single-cell transcriptomes via cFIT. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
57. Peng, M. *et al.* Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab481.
58. Polioudakis, D. *et al.* A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. *Neuron* **103**, 785–801.e8 (2019).
59. Carroll, L. S. & Owen, M. J. Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Med.* **1**, 102 (2009).
60. van der Sluijs, P. J. *et al.* The ARID1B spectrum in 143 patients: from nonsyndromic intellectual disability to Coffin-Siris syndrome. *Genet. Med.* **21**, 1295–1307 (2019).
61. Antaki, D. *et al.* A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. *medRxiv* (2021).
62. Wang, T. *et al.* Integrated gene analyses of de novo mutations from 46,612 trios with autism and developmental disorders. *bioRxiv* 2021.09.15.460398 (2021) doi:10.1101/2021.09.15.460398.

Methods

We confirm that this research complies with all relevant ethical regulations and was approved by the Mass General Brigham Human Research Committee (MGBHRC) Institutional Review Board (IRB) of Mass General Brigham:

Study Protocol 2012P001018, The Study of Novel Autism Genes and Other Neurodevelopmental Disorders (March 12, 2021)
Study Protocol 2013P000323, Genomic Studies of Human Neurodevelopment (September 07, 2018)

Protocols undergo annual continuing review by the MGBHRC IRB, Mass General Brigham, 399 Revolution Drive, Suite 710, Somerville, MA 02145. All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived. No participant compensation was provided from this study.

SNV/indel processing

ASD samples were aggregated from four independent sources: (1) previously published data from the Autism Sequencing Consortium (ASC; total N = 26,268[4,63]); (2) previously published data from the Simons Foundation Autism Research Initiative (SFARI) Simons Simplex Collection (SSC; total N = 9,170[63,64]); (3) unpublished data from the ASC (N = 5,036); and (4) the recently released Simons Foundation Powering Autism Research for Knowledge (SPARK initiative; N = 22,766[65]). The distribution of these samples is provided in **Supplementary Table 1** (one family is in both SPARK and unpublished ASC data, with different probands; one mother in the unpublished ASC data is also a proband in a different trio in the same dataset).

From these sources, family-based samples were processed and jointly genotyped in four batches. The first two batches included the published and unpublished ASC and SSC cohorts: 1) "ASC B14" included ASC samples through consortium sequencing batch 14 plus the SSC (N = 24,099, 4,632 new, 14,415 males and 9,684 females), and 2) "ASC B15-16" included ASC batches 15 and 16 (N = 832, all new, 500 males and 332 females). The latter two batches included two independent releases of the **SPARK** cohort: 3) the "SPARK Pilot" initial release (N = 1,379, 833 males and 546 females), and 4) the SPARK.27k.201909 ("SPARK main freeze") release (N = 21,387, 12,679 males and 8,708 males).

Raw sequencing outputs were aligned where needed to the GRCh38 reference genome and variants were jointly called following GATK[66] best practices. Briefly, individual gVCFs were generated by GATK HaplotypeCaller in gVCF mode and subsequently jointly genotyped for high confidence alleles using GenotypeGVCFs, accompanied by Variant Quality Score Recalibration (VQSR) to produce output VCFs. For additional details containing the specifics for each of the four batches, please see **Supplementary Note**. Finally, raw data was not available for 1,354 children and family members reported in Satterstrom *et al.*[4], and these variants were lifted over directly to GRCh38 (**Supplementary Table 1**).

SNV/indel filtering

Creation of working datasets

Hail 0.2 (https://hail.is) was used to process VCFs and write working datasets. Reported relationships and sample uniqueness were verified, sex was imputed, and variant consequences were annotated. Genotypes were filtered based on (1) depth, (2) genotype quality (GQ), (3) phred-scaled likelihood of the call being homozygous reference (PL[HomRef]),

(4) allele balance, (5) number of informative reads, and (6) Hardy-Weinberg p value. For additional details please see **Supplementary Note**.

De novo variant calling and quality control

For curation of *de novo* variants, we used Hail's de_novo() function to identify candidate variants, taking into account population variant frequencies. Candidates were further filtered based on: (1) frequency in gnomAD population and within their respective dataset, (2) "ExcessHet" filter, (3) allele balance and parent/child depth ratio, (4) VQSLOD values, and (5) excess number of *de novo* candidate variants within the same sample. For additional details please see **Supplementary Note**.

Case-control variants

ASC case-control samples consisted of Danish iPSYCH samples and Swedish PAGES samples. Rare variant counts for 4,863 autism and 5,002 control samples from the iPSYCH cohort were taken from the data of Satterstrom *et al.* (2019)[67], where rare variants were defined as those with an allele count no greater than 5 in the combination of the iPSYCH data with non-Finnish Europeans from the non-psychiatric subset of gnomAD (a total of 58,121 people). In addition to samples labeled as "Autism", samples labeled as "Both" in that study (meaning that an individual had both autism and ADHD diagnoses) were used as autism cases for our purposes. Rare variant counts for 728 autism and 3,595 control samples from the PAGES cohort were taken from Satterstrom *et al.* (2020)[4], where rare variants were defined as those with an allele count no greater than 5 in the 18,153 combined parents, cases, and controls in the dataset, as well as an allele count no greater than 5 in the non-psychiatric subset of ExAC r0.3 (45,376 people). Counts were removed for 17 cases for whom parental sequences became available, so that they are now included in our family-based data instead.

Transmitted variants

Counts of transmitted and non-transmitted alleles were produced starting from each of the four working datasets described above. First, variants were dropped that had been marked "ExcessHet" in the Filters field by GATK or had allele frequencies greater than 0.1% in either their own dataset or the non-neuro subset of gnomAD GRCh38 exomes v2.1.1. In addition, a filter requiring a GQ of at least 25 was applied to every genotype. Hail's transmission_disequilibrium_test() function was then called to count transmitted and untransmitted alleles for each variant in family-based data. Subsequently, additional dataset-specific filters on VQSLOD values were applied to derive final counts of transmitted and non-transmitted alleles. For additional details please see **Supplementary Note**.

CNV processing

For the subset of samples with available raw genomic data (**Supplementary Table 3**), we employed GATK-gCNV for exome CNV detection, along with an additional supplement of 7,832 general research use (GRU) controls. GATK-gCNV is a Bayesian method specifically designed to adjust for known bias factors in exome capture and sequencing (*e.g.* GC content), while automatically controlling for other technical and systematic differences. Briefly, raw sequencing files were compressed into read counts over the set of annotated exons and used as input, and a PCA-based approach was implemented on observed read counts to distinguish differences in capture kits (**Supplementary Fig. 1**), followed by a hybrid density- and distance-based clustering approach to curate batches of samples for parallel processing. After batching determination, GATK-gCNV was run for each batch and filtering metrics produced by the underlying Bayesian model were used to balance between sensitivity and positive predictive value (PPV). For details please see **Supplementary Note**. Of note, we observed five instances

among probands of possibly complex *de novo* SVs on chromosome 15, exhibiting adjacent GD duplications of differing copy states (**Supplementary Table 9**)

CNV Benchmarking

We had access to 8,439 samples for which matching genome and exome sequencing data were available for benchmarking comparisons. The ground truth data were CNVs called from WGS using the ensemble machine learning method GATK-SV[25,26]. After removing samples that did not pass GATK-gCNV exome QC filters (**Supplementary Note**, n=971 samples) and removing samples that had an outlier number of rare (site frequency <1%) calls in the GATK-SV genome callset (>16 rare calls, based on median+2*interquartile range, n=477 samples), 7,035 samples remained for direct comparison. Benchmarking was carried out for all rare CNVs (site frequency < 1%). Sensitivity was measured by the proportion of sites called from WGS data that had a match in the GATK-gCNV callset. Specifically, for each site, if at least 50% of the samples that had that CNV in the WGS data also had a GATK-gCNV call with a consistent direction (deletion or duplication) that overlapped at least 50% of the captured intervals, this was considered a success. For CNVs called by GATK-gCNV, PPV was measured by requiring that 50% of the GATK-gCNV samples with that call had a match in the WGS calls (ground truth) with at least 50% interval overlap. We evaluated sensitivity and PPV as a function of the number of captured exons overlapping the canonical transcripts of protein-coding genes.

TADA Bayesian Framework for Gene Association

TADA is a Bayesian framework that produces gene-level measures of evidence for association that can be transformed into a false discovery rate[42]. Broadly speaking, for a given variant type and gene, TADA produces a Bayes Factor (BF) to measure statistical evidence, taking as input the count of variant events, the mutation rate, the number of samples, and a prior on the risk of a variant in each gene. BF can be readily combined across different variant types for the same gene by multiplication, arriving at a total measure of association for a given gene. This total BF can then be directly transformed into a FDR and the appropriate statistical threshold can be applied to extract a candidate gene list. In the previous TADA study[4], evidence was aggregated for *de novo* PTVs, MisB variants, and MisA variants, as well as case/control PTVs to find 102 genes meeting an FDR ≤ 0.1 threshold.

For this analysis, we have extended TADA to leverage updated measures of constraint (LOEUF), the full combination of *de novo*, case/control, and inherited x PTV, MisB, MisA, deletion, and duplication variants, as well as variants in unaffected siblings. For full details please see **Supplementary Note**.

Applying TADA to DD data

We accessed the summary tables released by the DDD in Kaplanis *et al*.[5], detailing *de novo* variants detected and gene-level variant counts in 31,058 trios where the offspring was diagnosed with a developmental disorder. To calculate the number of PTVs per gene, we aggregated the Kaplanis *et al.* variants annotated with consequences of "frameshift_variant", "splice_donor_variant", "splice_acceptor_variant", or "stop_gained". For synonymous counts, we aggregated variants with labels of “synonymous_variant” or “stop_retained_variant”. We annotated missense variants (“missense_variant”) with MPC scores, and using those MPC scores, we assigned MisB and MisA status and aggregated counts per gene.

To create TADA-DD, we supplied the per-gene counts of PTVs, MisA variants, and MisB variants to TADA in the same manner as we supplied our ASD cohort counts. TADA-DD BFs

were then combined with those from the ASD cohort on a per-gene basis, allowing us to estimate FDR on a combined NDD super-cohort (TADA-NDD).

Comparison of TADA-DD and denovoWEST from Kaplanis *et al.*

Kaplanis *et al.* reports association values for 19,654 genes, of which 285 are significant at an exome-wide threshold. Of the 18,128 autosomal genes investigated by our study, 17,919 (99%) have a match from Kaplanis *et al,* including all 252 significant autosomal genes. Of the Kaplanis *et al.* denovoWEST exome-wide significant genes, 237/252 (94%) also appear in the TADA-DD FDR ≤ 0.001 list.

We also measured the concordance of the Bayesian TADA-DD FDR with the frequentist denovoWEST estimates of gene significance reported in Kaplanis *et al.* by transforming the Kaplanis p-values (denovoWEST_p_full) into FDRs (FDR denovoWEST) using the R function p.adjust(method="fdr"). A pairwise plot of TADA-DD FDR with transformed Kaplanis FDR reveals high concordance (**Supplementary Fig. 6**, cor=0.95) on the log scale, signaling convergence in evaluation of gene-level evidence between our studies, and allowing us to integrate the Kaplanis variant data in our Bayesian framework.

Female protective effect versus ascertainment bias of affected females

Severity of phenotype and sex are known to be associated with the presence of *de novo* SNV/indel or CNV mutations in individuals with ASD. Specifically, those with more severe phenotypes or females are more likely to be carriers of such mutations. Notably, various studies have also found that females are less likely to be diagnosed with ASD compared to males with similar presentation[51,52,68], creating the possibility that the excess burden of damaging *de novo* variants observed in females could be due to this ascertainment bias – females are simply more severely affected. An alternative is that severity and sex combine approximately additively to determine burden. This would be consistent with an alternative hypothesis, a female protective effect, which posits that females require a greater burden of genetic risk variation to be affected.Using ADOS and IQ as proxies for severity where available, we constructed logistic regression models of carrier status as the outcome and sex, severity, and their interactions as predictors. We found no evidence to support ascertainment bias and instead favor the additive alternative (**Supplementary Table 13**). For more details please see **Supplementary Note**.

Evaluating oligogenicity in ASD and DD

We tabulated the number of individuals with 0, 1, and 2 *de novo* damaging variants (PTV or MisB) among the TADA-ASD 72, TADA-ASD 185, and TADA-NDD 373 genes and constructed a poisson expectation on the number of expected individuals with 2 such variants as follows:

p = (number of variants/number of samples)^2
Expectation = (p * number of samples) * exp(-p)/2!

These analyses also offer a glimpse into the evidence supporting an oligogenic model of ASD and DD. Using the list of 373 NDD-associated genes, we observed 913 (6.1%) of the 15,036 ASD probands harboring a damaging *de novo* variant of interest (PTV or MisB), and 12 probands that carried two (0.08%). Across all 31,058 DD probands, one *de novo* variant was found in 5,176 (16.7%) cases, and 96 (0.31%) carried two. Using a Poisson expectation model for the number of affected individuals carrying two variants, we find depletion in both the ASD and DD cases carrying two variants (ASD: 27.4 expected, 12 observed; DD: 390 expected, 96 observed). This same depletion was observed when restricting to the 72 or 185 genes associated with ASD alone, indicating no support for oligogenicity among ASD or DD cases from these analyses (**Supplementary Tables 14a-c**).

Conditional analysis of cross-cohort association

For the ASD and DD cohorts, separately, we first converted the set of 18,128 gene q-values into p-values using the following R command: pval = qval * rank(qval) / (max(qval) * length(qval)). Next, we selected genes meeting FDR ≤ 0.05 from the TADA-ASD and TADA-DD cohorts, treating the derived lists separately. For the set of 185 identified TADA-ASD genes, we evaluated the distribution of their back-transformed p-values from TADA-DD using the 'propTrueNull' function from the R package 'limma_3.38.3'[69,70] to estimate pi0 and pi1=1-pi0, which is the estimated fraction of the number of genes associated in the DD cohort. pi0 is the estimated fraction of genes that have no association and for which their p-values would be uniformly distributed on the interval 0-1. We then did the converse: choosing the set of 477 identified TADA-DD genes, we evaluated the distribution of their p-values from TADA-ASD to estimate pi1.

ASD-DDD heterogeneity analysis

We asked which of the 464 signal genes was more tightly connected with either ASD or DD than expected by chance. To do so, we formulated an approach that builds on the familiar chi-statistic residual. Before computing the residuals, we needed to overcome the far larger number of mutations present in the DD sample because the standardized residual performs best when the total count of events, per cohort, is equal. We therefore down-sampled the DD mutations in signal genes to obtain a count of 1001 mutations, matching the count of mutations in the ASD cohort. This was repeated for 100 repetitions.

C statistic. For the C statistic, we used a standard log-linear model analysis by conditioning on the row (gene) and column totals (over ASD or DDD). We asked if the residual for ASD was substantially different from that expected under the null. The residual for gene i was defined as

$$C_i^{ds} = \left(dn.asd_i^{obs} - dn.asd_i^{exp}\right) / \sqrt{dn.asd_i^{exp}}$$

where:

$$dn.asd_i^{exp} = dn.asd_i^{obs} \times \left(dn.asd_i^{obs} + dn.ddd_i^{obs}\right)/N$$

with the average over the 100 down-sampling repetitions was recorded as the C statistic for each gene.

Mixture modeling

If gene mutation rates were independent of cohort, then the C statistic would be distributed as a standard normal statistic, but this was clearly not true (**Fig. 5e**). Genes with unusually few mutations in the ASD cohort produced a negative C statistic and those with unusually many mutations in the ASD cohort produced a positive statistic. Assuming the genes split into two classes, one favoring DD mutations and the other favoring ASD mutations, we fitted a two-component normal mixture model. This calculation was performed using the normalmixEM function in the R library[71]. We restricted the model to have a common standard deviation for both components (option arbvar=F), which was estimated to be 0.527. Although the C statistics varied continuously across the spectrum of values observed, we could estimate the posterior probability a gene was from the DD or ASD component to determine likely group membership. Genes with posterior probability greater than 0.99 for either class were labeled by that class.

Tree analysis

To understand the developmental cell types in which these genes were expressed, we analyzed two datasets using a new approach called cFIT, the Common Factor Integration and Transfer

learning algorithm[56]. cFIT relies on a linear model assuming a common factor matrix shared among datasets, as well as gene-wise location and scale shifts unique to each dataset. It estimates the shared and batch specific parameters through iterative nonnegative matrix factorization and then recovers the batch-free expression for each dataset based on the common factor and factor loadings. We applied cFIT to fetal cells from two studies[55,58] and used unsupervised clustering (MRtree)[57] to the integrated data to generate a hierarchical tree of various cell types. For details please see **Supplementary Note**.

Enrichment analysis

We performed enrichment analysis for each cluster in the resulting tree to determine if any clusters expressed an unusual number of ASD-predominant or DD-predominant risk genes. Before performing the enrichment analysis, i.e., creating a 2x2 table for expressed gene (yes/no) by risk gene (yes/no), we needed to first identify the set of genes to be included in the analysis, which is defined as the set of genes "expressed" in at least one cell type. Because the integration process often replaces zero values in the gene expression matrix with small positive values, we considered any integrated expression value less than 0.5 to be non-expressed. A gene was considered "expressed" for a particular cell type if its expression was greater than 0.5 for at least 25% of the cells in the terminal clusters. For each cluster, we then determined if the expressed genes belonged to the ASD-predominant or DD-predominant gene sets and computed the odds ratio from the 2x2 table to determine enrichment (**Fig. 5c**).

Evaluating overlap with Schizophrenia-associated genes

We compared our ASD- and DD-associated genes to the schizophrenia-associated genes reported by SCHEMA[7] to determine if there was any overlap between ASD and schizophrenia at the level of individual risk genes and, if so, whether it was related to ASD-DD overlap. Note that 3/309 genes with FDR ≤ 0.001 in TADA-DD were not included in the SCHEMA results, while 10/244 genes identified by SCHEMA as schizophrenia-associated at $p < 0.01$ were not evaluated by our TADA model (8/10 were on chr X).

As described in the main text, among the 72 ASD genes we discovered at an FDR ≤ 0.001, 61 show an association with DD (using FDR ≤ 0.001, based on TADA-DD), and 8 show an association with schizophrenia at $p < 0.01$. If the two associations were independent, we would expect ~1 of the 8 ASD-schizophrenia genes to lack an association with DD (based on all but 11/72 = ~15% of the ASD genes overlapping DD). However, we in fact find that 4 of the ASD-schizophrenia genes lack an association with DD, which is a significant overrepresentation compared to random chance ($p = 0.023$, binomial test, **Supplementary Fig. 11a)**.

We also analyzed ASD-schizophrenia overlap using the 36 ASD-predominant genes and 82 DD-predominant genes shown at the extremities of the distribution in **Fig. 5f** (which shows posterior probability for ASD enrichment of the genes in our heterogeneity analysis). We looked for overlap between these gene sets and the 244 genes identified by SCHEMA as schizophrenia-associated at $p < 0.01$. We found that 6 of the 36 ASD-predominant genes were schizophrenia-associated, while 3 of the 82 DD-predominant genes were schizophrenia-associated (**Supplementary Fig. 11b**). If we compare to the null hypothesis that each of the 17,294 genes from our TADA model that are also in the SCHEMA results has an equal chance of being schizophrenia-associated, then the ASD-schizophrenia overlap is significantly enriched ($p = 8.4 \times 10^{-6}$, binomial test), while the DD-schizophrenia overlap is not ($p = 0.10$, binomial test). The two outcomes (6/36 vs. 3/82) are also different when compared to each other ($p = 0.023$, Fisher's exact test).

Data Availability

The data used in this study are available at:
Repository/DataBank Accession: NHGRI AnVIL
Accession ID: phs000298
Databank URL: https://anvilproject.org/data

Repository/DataBank Accession: Simons Foundation for Autism Research Initiative SFARIbase
Accession ID: SPARK/Regeneron/SPARK_WES_2/
Databank URL: https://www.sfari.org/resource/spark/

De novo variant data used analyses are reported in **Supplementary Table 9** (CNVs) and **Supplementary Table 20** (SNV/indels). Other candidate *de novo* CNVs that were either too small (spanning two exons or less) or did not meet QS threshold (QS<200) to be included in our statistical analyses are reported in **Supplementary Table 21**. Aggregated rare variant counts (inherited, case/control) are released in **Supplementary Tables 5-7**. To access all individual variants, please see above repositories.

GRCh38 reference genome: gs://gcp-public-data--broad-references/hg38/v0/Homo_sapiens_assembly38.fasta

Access to UK Biobank data will be provided by the UK Biobank.

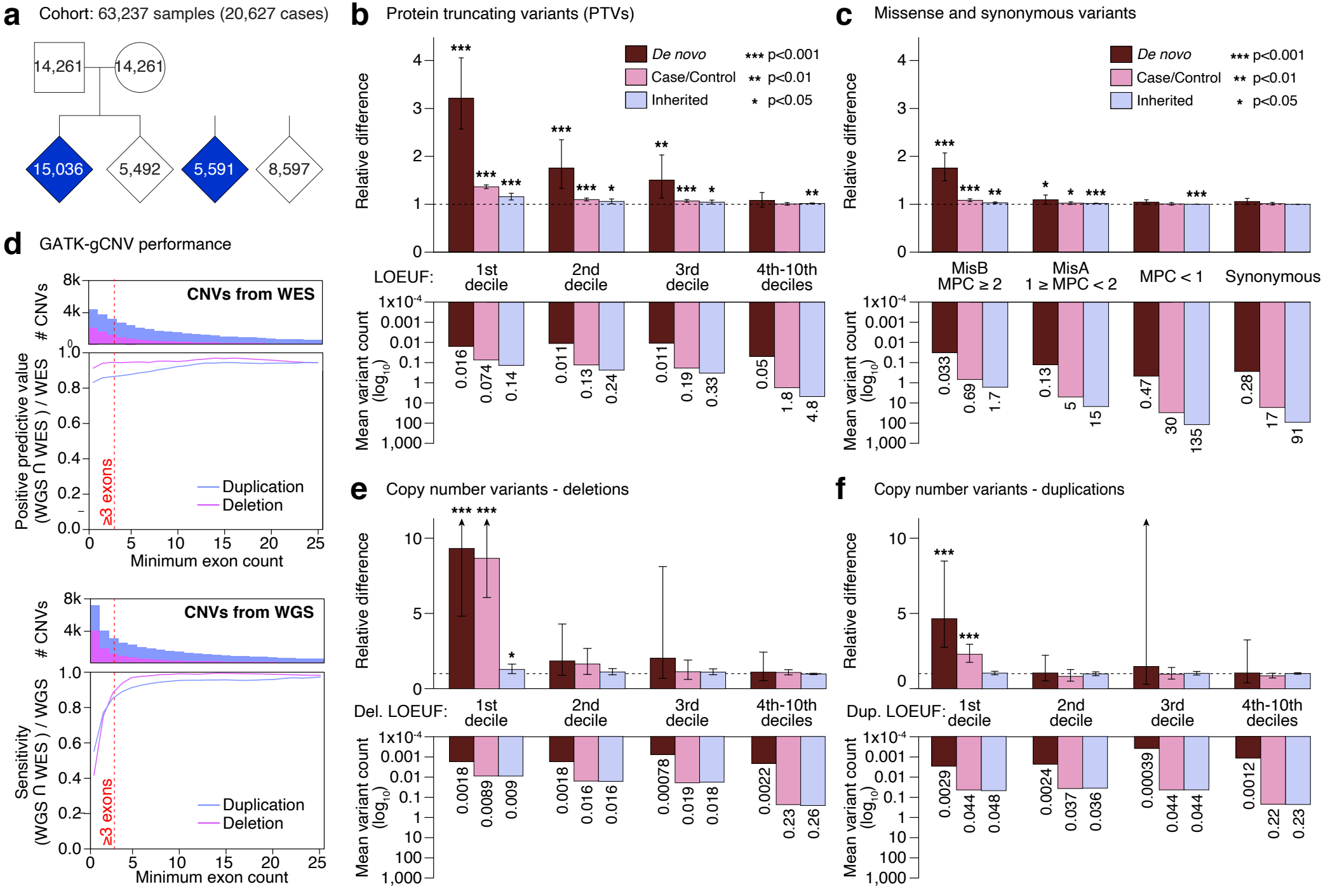
Code Availability

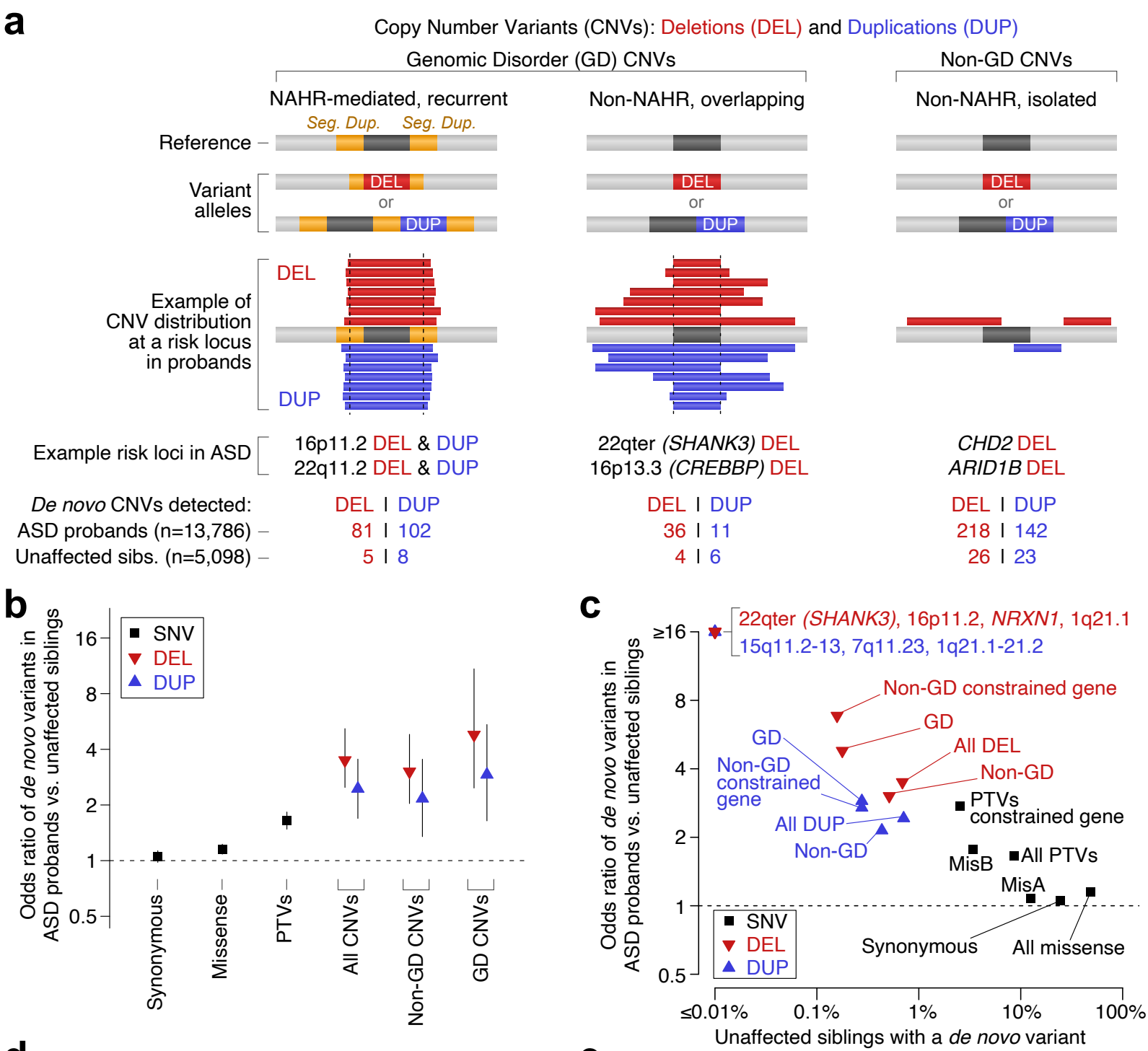
The R code used to generate TADA association results is available under the MIT license at https://github.com/talkowski-lab/TADA_2022
DOI:10.5281/zenodo.6496480

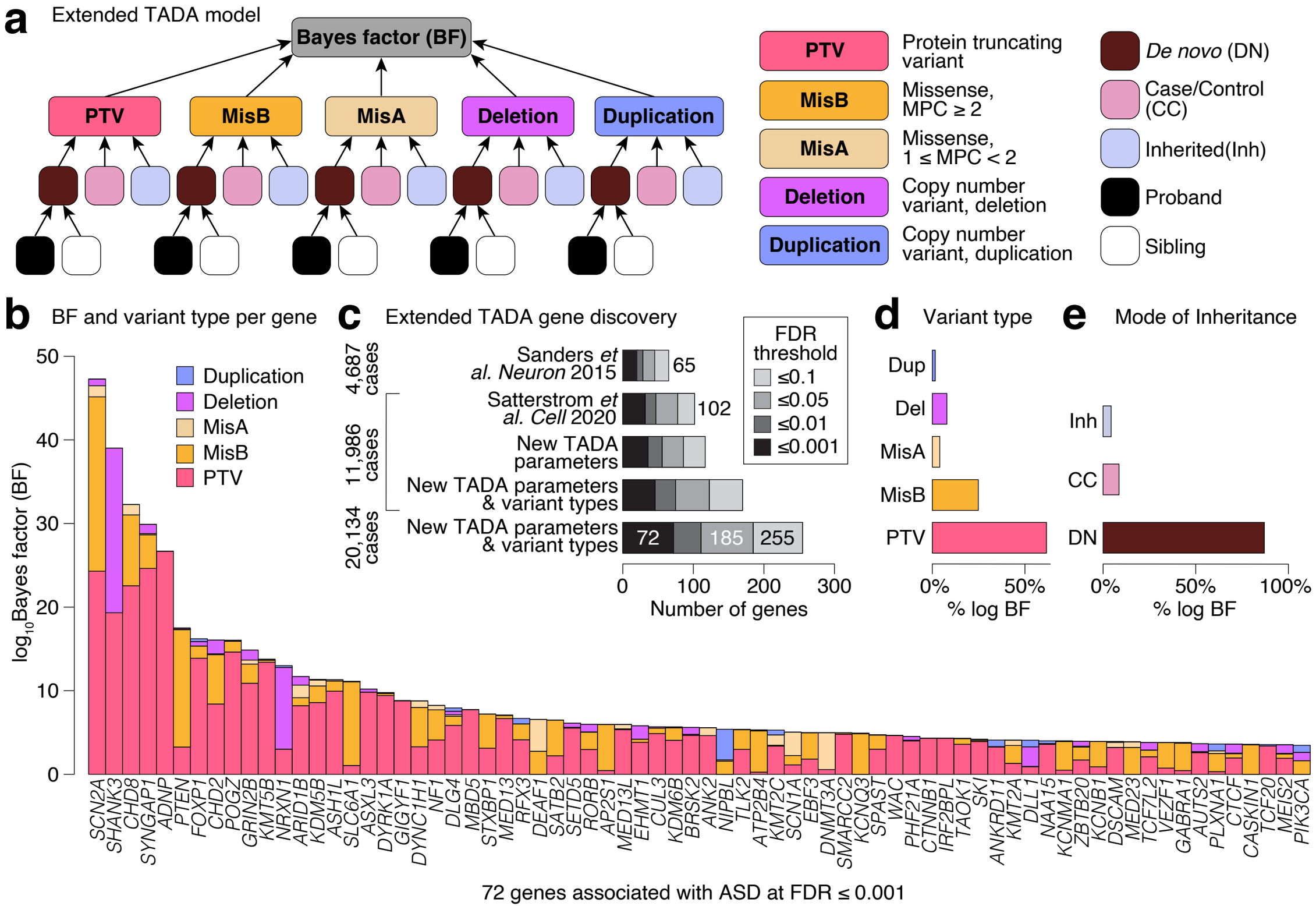
Analyses executed in R 3.5.3: limma_3.38.3, stringr_1.4.0, GenomicRanges_1.34.0, GenomeInfoDb_1.18.1, IRanges_2.16.0, S4Vectors_0.20.1, BiocGenerics_0.28.0.

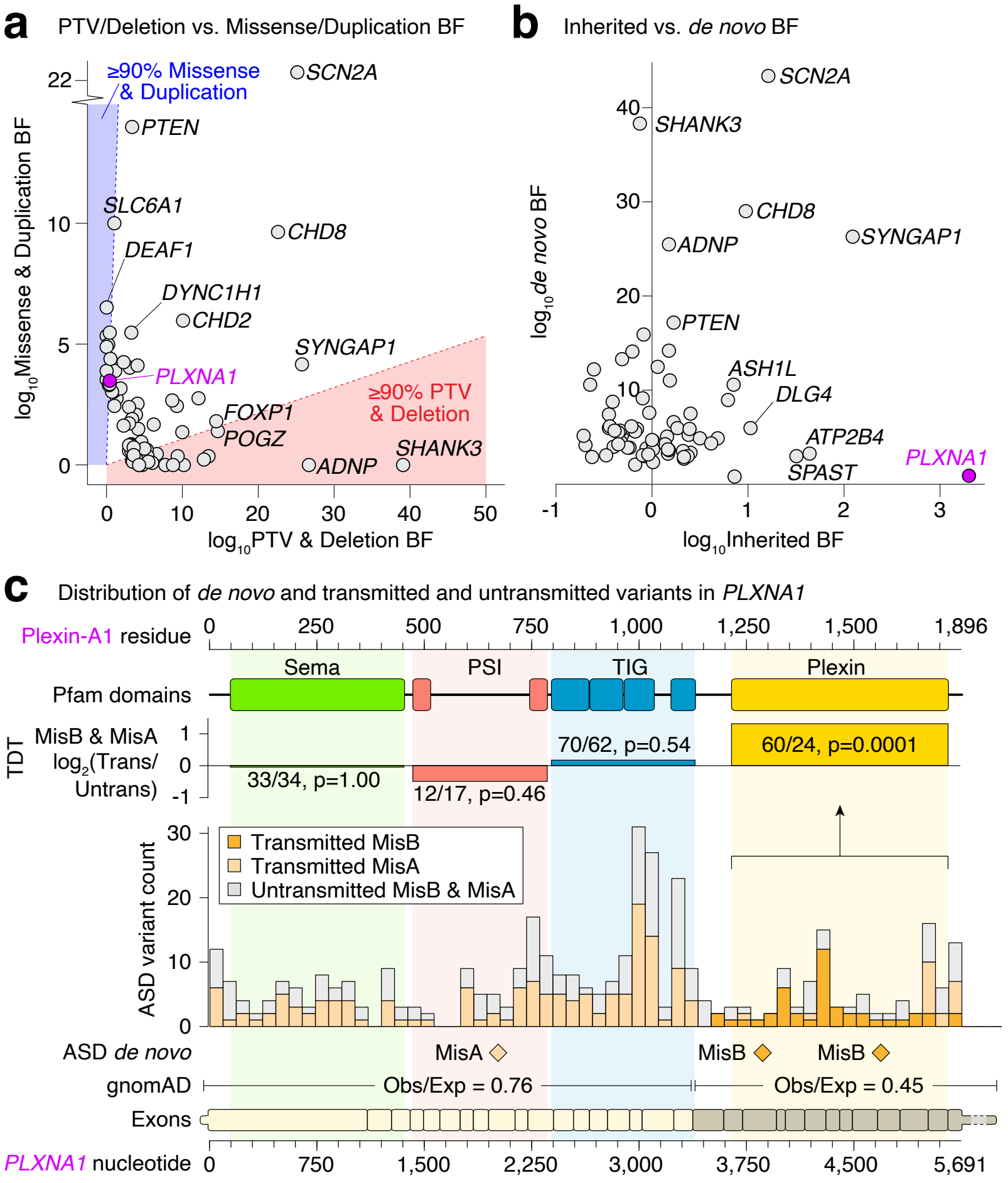
Methods-only References

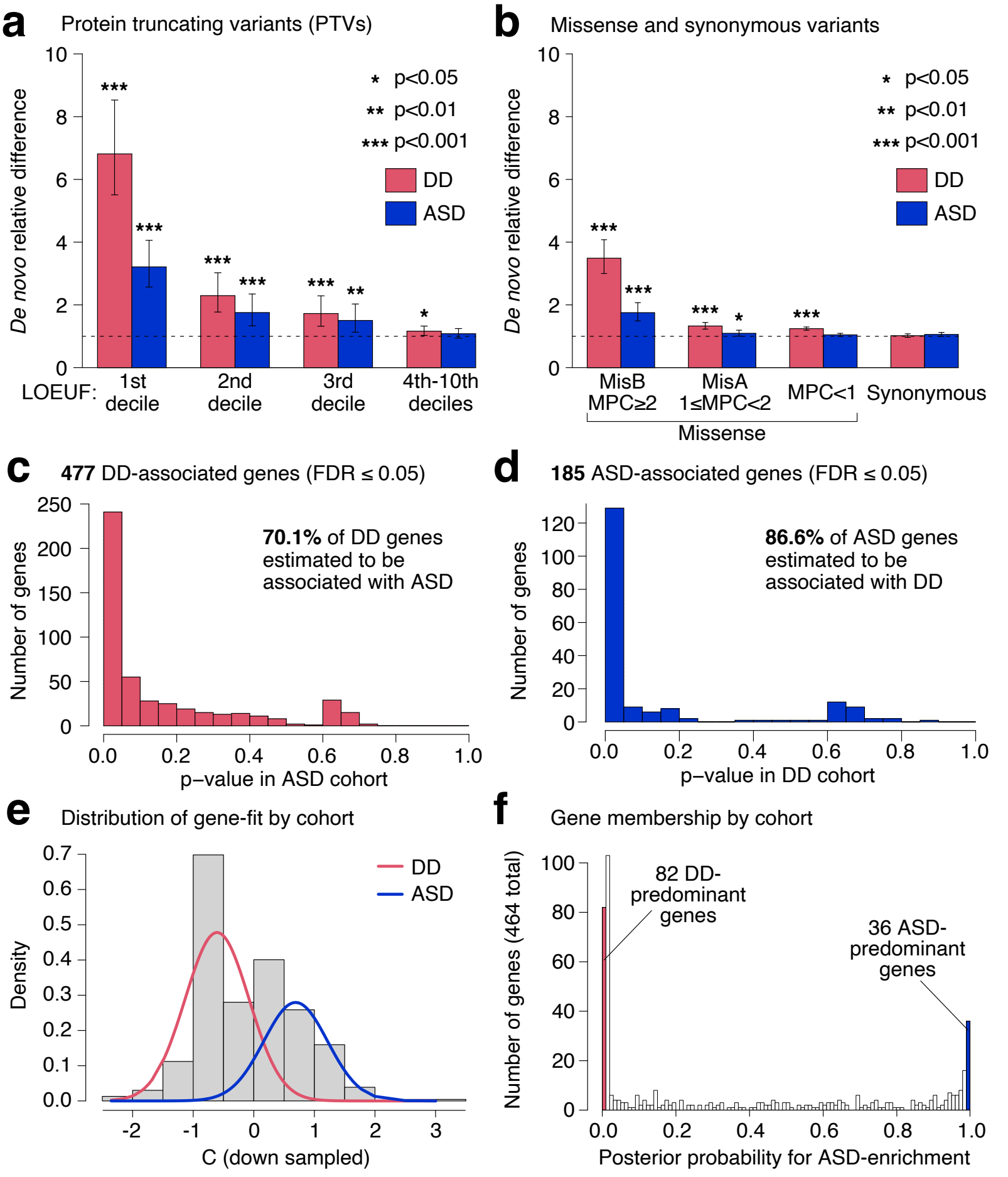
63. Buxbaum, J. D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052–1056 (2012).
64. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
65. SPARK Consortium. Electronic address: pfeliciano@simonsfoundation.org & SPARK Consortium. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018).
66. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. ('O'Reilly Media, Inc.', 2020).
67. Satterstrom, F. K. *et al.* Autism spectrum disorder and attention deficit hyperactivity disorder have a similar burden of rare protein-truncating variants. *Nat. Neurosci.* **22**, 1961–1965 (2019).
68. Loomes, R., Hull, L. & Mandy, W. P. L. What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis. *J. Am. Acad. Child Adolesc. Psychiatry* **56**, 466–474 (2017).
69. Jiang, H. & Doerge, R. W. Estimating the proportion of true null hypotheses for multiple comparisons. *Cancer Inform.* **6**, 25–32 (2008).
70. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
71. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: AnRPackage for Analyzing Finite Mixture Models. *Journal of Statistical Software* vol. 32 (2009).

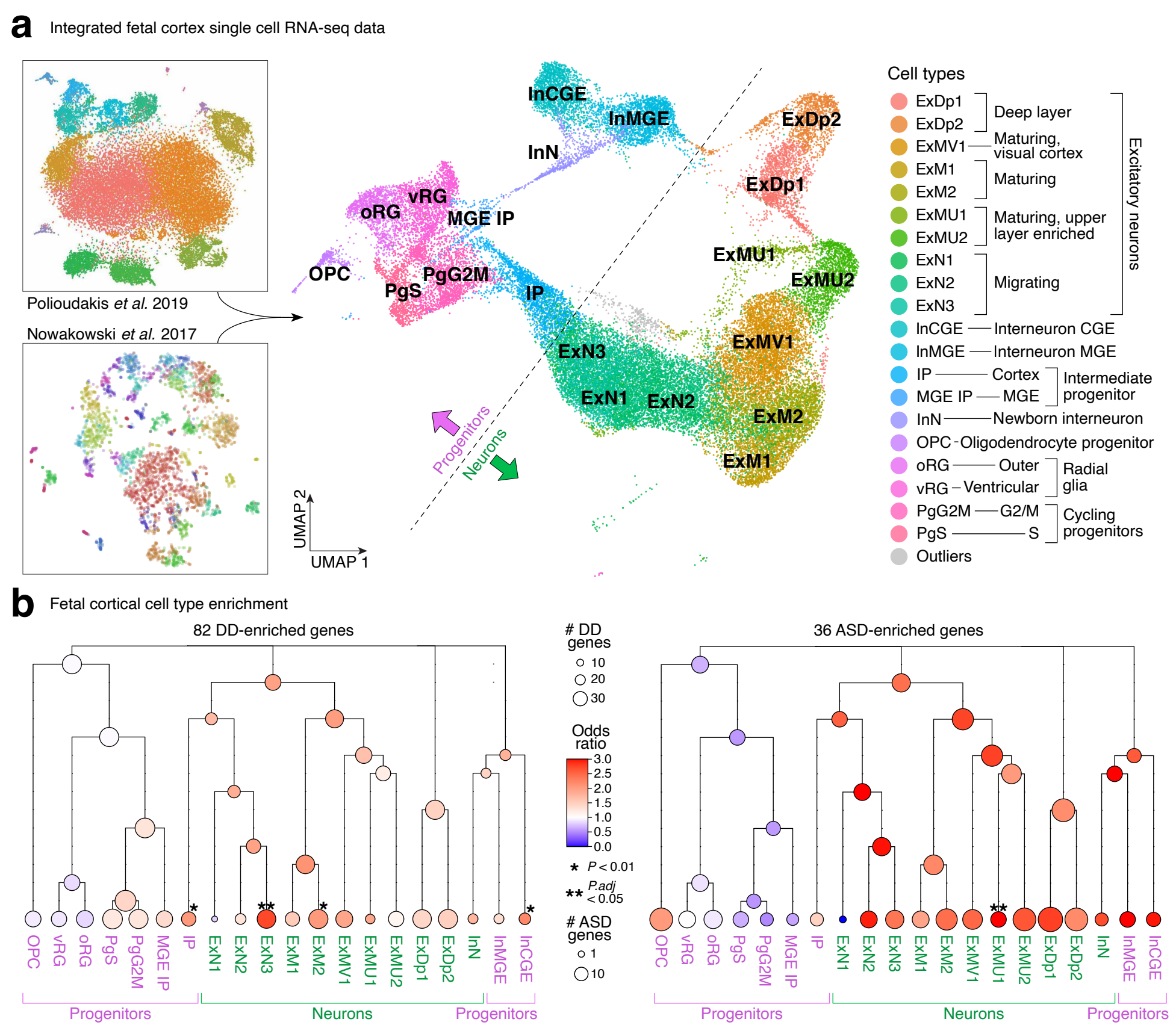

a
Cohort: 63,237 samples (20,627 cases)
14,261
14,261
15,036
5,492
5,591
8,597
b
Protein truncating variants (PTVs)
c
Missense and synonymous variants
De novo
Case/Control
Inherited
*** p<0.001
** p<0.01
* p<0.05
Relative difference
LOEUF:
1st decile
2nd decile
3rd decile
4th-10th deciles
MisB MPC ≥ 2
MisA 1 ≥ MPC < 2
MPC < 1
Synonymous
Mean variant count (log10)
0.016
0.074
0.14
0.011
0.13
0.24
0.011
0.19
0.33
0.05
1.8
4.8
0.033
0.69
1.7
0.13
5
15
0.47
30
135
0.28
17
91
d
GATK-gCNV performance
CNVs from WES
CNVs
Positive predictive value (WGS ∩ WES) / WES
Duplication
Deletion
≥3 exons
Minimum exon count
CNVs from WGS
Sensitivity (WGS ∩ WES) / WGS
e
Copy number variants - deletions
Del. LOEUF:
0.0018
0.0089
0.009
0.0018
0.016
0.016
0.00078
0.019
0.018
0.0022
0.23
0.26
f
Copy number variants - duplications
Dup. LOEUF:
0.0029
0.044
0.048
0.0024
0.037
0.036
0.00039
0.044
0.044
0.0012
0.22
0.23


a
Copy Number Variants (CNVs): Deletions (DEL) and Duplications (DUP)
Genomic Disorder (GD) CNVs
Non-GD CNVs
NAHR-mediated, recurrent
Non-NAHR, overlapping
Non-NAHR, isolated
Seg. Dup.
Seg. Dup.
Reference
Variant alleles
DEL
or
DUP
Example of CNV distribution at a risk locus in probands
Example risk loci in ASD
16p11.2 DEL & DUP
22q11.2 DEL & DUP
22qter (SHANK3) DEL
16p13.3 (CREBBP) DEL
CHD2 DEL
ARID1B DEL
De novo CNVs detected:
DEL | DUP
ASD probands (n=13,786)
81 | 102
36 | 11
218 | 142
Unaffected sibs. (n=5,098)
5 | 8
4 | 6
26 | 23
b
Odds ratio of de novo variants in ASD probands vs. unaffected siblings
SNV
DEL
DUP
Synonymous
Missense
PTVs
All CNVs
Non-GD CNVs
GD CNVs
c
22qter (SHANK3), 16p11.2, NRXN1, 1q21.1
15q11.2-13, 7q11.23, 1q21.1-21.2
Non-GD constrained gene
GD
All DEL
Non-GD
PTVs constrained gene
All DUP
Non-GD
MisB
All PTVs
MisA
Synonymous
All missense
Unaffected siblings with a de novo variant
d
Odds ratio in ASD probands vs. UK Biobank controls
15q11.2–13.1
17q12
17q11.2
7q11.23
22q11.21
22q13.2–q13.33
15q13.1–13.2
16p11.2
16p11.2
15q13.2–q13.3
1q21.1–q21.2
16p11.2 distal
16p13.11
22q13.2–q13.33
Frequency in UK Biobank
e
CNV Criteria
Phenotype
Maternal
Paternal
All de novo
Unaffected
All de novo
ASD
De novo Non-NAHR GDs
ASD
De novo NAHR GDs
ASD
De novo 16p11.2 GD
ASD
Proportion with maternal origin


a
Extended TADA model
Bayes factor (BF)
PTV
MisB
MisA
Deletion
Duplication
PTV
Protein truncating variant
MisB
Missense, MPC ≥ 2
MisA
Missense, 1 ≤ MPC < 2
Deletion
Copy number variant, deletion
Duplication
Copy number variant, duplication
De novo (DN)
Case/Control (CC)
Inherited(Inh)
Proband
Sibling
b
BF and variant type per gene
Duplication
Deletion
MisA
MisB
PTV
log10Bayes factor (BF)
50
40
30
20
10
0
c
Extended TADA gene discovery
4,687 cases
11,986 cases
20,134 cases
Sanders et al. Neuron 2015
65
Satterstrom et al. Cell 2020
102
New TADA parameters
New TADA parameters & variant types
New TADA parameters & variant types
72
185
255
FDR threshold
≤0.1
≤0.05
≤0.01
≤0.001
0
100
200
300
Number of genes
d
Variant type
Dup
Del
MisA
MisB
PTV
0%
50%
% log BF
e
Mode of Inheritance
Inh
CC
DN
0%
50%
100%
% log BF
SCN2A
SHANK3
CHD8
SYNGAP1
ADNP
PTEN
FOXP1
CHD2
POGZ
GRIN2B
KMT5B
NRXN1
ARID1B
KDM5B
ASH1L
SLC6A1
ASXL3
DYRK1A
GIGYF1
DYNC1H1
NF1
DLG4
MBD5
STXBP1
MED13
RFX3
DEAF1
SATB2
SETD5
RORB
AP2S1
MED13L
EHMT1
CUL3
KDM6B
BRSK2
ANK2
NIPBL
TLK2
ATP2B4
KMT2C
SCN1A
EBF3
DNMT3A
SMARCC2
KCNQ3
SPAST
WAC
PHF21A
CTNNB1
IRF2BPL
TAOK1
SKI
ANKRD11
KMT2A
DLL1
NAA15
KCNMA1
ZBTB20
KCNB1
DSCAM
MED23
TCF7L2
VEZF1
GABRA1
AUTS2
PLXNA1
CTCF
CASKIN1
TCF20
MEIS2
PIK3CA
72 genes associated with ASD at FDR ≤ 0.001


a
PTV/Deletion vs. Missense/Duplication BF
≥90% Missense & Duplication
SCN2A
PTEN
SLC6A1
CHD8
DEAF1
DYNC1H1
CHD2
SYNGAP1
PLXNA1
≥90% PTV & Deletion
FOXP1
POGZ
SHANK3
ADNP
log10Missense & Duplication BF
log10PTV & Deletion BF
b
Inherited vs. de novo BF
SCN2A
SHANK3
CHD8
SYNGAP1
ADNP
PTEN
ASH1L
DLG4
ATP2B4
SPAST
PLXNA1
log10de novo BF
log10Inherited BF
c
Distribution of de novo and transmitted and untransmitted variants in PLXNA1
Plexin-A1 residue
Sema
PSI
TIG
Plexin
Pfam domains
TDT
MisB & MisA log2(Trans/ Untrans)
33/34, p=1.00
12/17, p=0.46
70/62, p=0.54
60/24, p=0.0001
Transmitted MisB
Transmitted MisA
Untransmitted MisB & MisA
ASD variant count
ASD de novo
MisA
MisB
MisB
gnomAD
Obs/Exp = 0.76
Obs/Exp = 0.45
Exons
PLXNA1 nucleotide


a
Protein truncating variants (PTVs)
De novo relative difference
* p<0.05
** p<0.01
*** p<0.001
DD
ASD
LOEUF:
1st decile
2nd decile
3rd decile
4th-10th deciles
b
Missense and synonymous variants
MisB MPC≥2
MisA 1≤MPC<2
MPC<1
Synonymous
Missense
c
477 DD-associated genes (FDR ≤ 0.05)
70.1% of DD genes estimated to be associated with ASD
Number of genes
p–value in ASD cohort
d
185 ASD-associated genes (FDR ≤ 0.05)
86.6% of ASD genes estimated to be associated with DD
p–value in DD cohort
e
Distribution of gene-fit by cohort
Density
C (down sampled)
f
Gene membership by cohort
Number of genes (464 total)
82 DD-predominant genes
36 ASD-predominant genes
Posterior probability for ASD-enrichment


a
Integrated fetal cortex single cell RNA-seq data
Polioudakis et al. 2019
Nowakowski et al. 2017
InCGE
InMGE
InN
vRG
oRG
MGE IP
OPC
PgS
PgG2M
IP
ExDp2
ExDp1
ExMU1
ExMU2
ExN3
ExMV1
ExN1
ExN2
ExM2
ExM1
Progenitors
Neurons
UMAP 2
UMAP 1
Cell types
ExDp1
ExDp2
ExMV1
ExM1
ExM2
ExMU1
ExMU2
ExN1
ExN2
ExN3
InCGE
InMGE
IP
MGE IP
InN
OPC
oRG
vRG
PgG2M
PgS
Outliers
Deep layer
Maturing, visual cortex
Maturing
Maturing, upper layer enriched
Migrating
Excitatory neurons
Interneuron CGE
Interneuron MGE
Cortex
MGE
Intermediate progenitor
Newborn interneuron
Oligodendrocyte progenitor
Outer
Ventricular
Radial glia
G2/M
S
Cycling progenitors
b
Fetal cortical cell type enrichment
82 DD-enriched genes
36 ASD-enriched genes
DD genes
10
20
30
Odds ratio
3.0
2.5
2.0
1.5
1.0
0.5
0.0
* P < 0.01
** P.adj < 0.05
# ASD genes
1
10
OPC
vRG
oRG
PgS
PgG2M
MGE IP
IP
ExN1
ExN2
ExN3
ExM1
ExM2
ExMV1
ExMU1
ExMU2
ExDp1
ExDp2
InN
InMGE
InCGE
Progenitors
Neurons
Progenitors