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

ABSTRACT

Individuals with autism spectrum disorder (ASD) or related neurodevelopmental disorders (NDDs) often carry disruptive mutations in genes that are depleted of functional variation in the broader population. We build upon this observation and exome sequencing from 154,842 individuals to explore the allelic diversity of rare protein-coding variation contributing risk for ASD and related NDDs. Using an integrative statistical model, we jointly analyzed rare protein-truncating variants (PTVs), damaging missense variants, and copy number variants (CNVs) derived from exome sequencing of 63,237 individuals from ASD cohorts. We discovered 71 genes associated with ASD at a false discovery rate (FDR) ≤ 0.001, a threshold approximately equivalent to exome-wide significance, and 183 genes at FDR ≤ 0.05. Associations were predominantly driven by de novo PTVs, damaging missense variants, and CNVs: 57.4%, 21.2%, and 8.32% of evidence, respectively. Though fewer in number, CNVs conferred greater relative risk than PTVs, and repeat-mediated de novo CNVs exhibited strong maternal bias in parent-of-origin (e.g., 92.3% of 16p11.2 CNVs), whereas all other CNVs showed a paternal bias. To explore how genes associated with ASD and NDD overlap or differ, we analyzed our ASD cohort alongside a developmental delay (DD) cohort from the deciphering developmental disorders study (DDD; n=91,605 samples). We first reanalyzed the DDD dataset using the same models as the ASD cohorts, then performed joint analyses of both cohorts and identified 373 genes contributing to NDD risk at FDR ≤ 0.001 and 662 NDD risk genes at FDR ≤ 0.05. Of these NDD risk genes, 54 genes (125 genes at FDR ≤ 0.05) were unique to the joint analyses and not significant in either cohort alone. Our results confirm overlap of most ASD and DD risk genes, although many differ significantly in frequency of mutation. Analyses of single-cell transcriptome datasets showed that genes associated predominantly with DD were strongly enriched for earlier neurodevelopmental cell types, whereas genes displaying stronger evidence for association in ASD cohorts were more enriched for maturing neurons. The ASD risk genes were also enriched for genes associated with schizophrenia from a separate rare coding variant analysis of 121,570 individuals, emphasizing that these neuropsychiatric disorders share common pathways to risk.

What remains less clear is the relative burden of CNVs at the resolution of individual genes or exons in ASD etiology, as studies to date have been limited by the relatively small samples assessed by whole-genome sequencing (WGS) and the high false positive rates for existing exome-based CNV discovery methods. Prior studies by the Exome Aggregation Consortium (ExAC), 100 Genomes Project, and Genome Aggregation Database (gnomAD) have established size distributions and mutation rates for structural variants (SVs) across human populations, while studies leveraging long-read WGS have provided further insights into the broader swath of human SVs, over 99% of which are not readily detectable by microarray. Here, we demonstrate that significant improvements in computational models to normalize exome sequencing read depth, combined with appropriate quality control filters, enable discovery of even small rare coding CNVs with validation rates comparable to indels. These methods allow us to build upon prior studies to now jointly analyze rare coding SNVs, indels, and CNVs at the resolution of individual genes and exons from large-scale ASD datasets.

Rare variant discovery can also be enhanced through the integration of functional data, particularly in cohorts such as ASD that display strong selective pressure and reduced fecundity. One such measure is the recently developed loss-of-function observed/expected upper bound fraction (LOEF) score, which is a continuous measure of selective pressure against PTVs in each gene. Similarly, the ‘missense badness’, PolyPhen-2, and constraint (MPC) score is one of several measures of the estimated deleteriousness of missense variation. In this study, we used a Bayesian statistical framework, the Transmission and De Novo Association (TADA) model, to incorporate these functional annotations and jointly analyze 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). Our analyses 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 cell types in which they are enriched during early neuronal development. 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.

Background

Autism spectrum disorder (ASD) affects over 1.7% of children in the United States. Epidemiological studies have repeatedly demonstrated that ASD is highly heritable, with the majority of risk stemming from common genetic variants, each of small effect, acting additively across the genome. In ASD and related NDDs, many rare deleterious variation studies have noted modest enrichment of rare inherited variants as well. Exome sequencing provides an efficient method to detect these rare and de novo variants, which has led to the association of numerous genes with ASD through repeated observations of such variants across independent studies. Beyond ASD, exome sequencing has enabled the discovery of genes associated with overlapping and distinct genetic architectures across a spectrum of developmental and neuropsychiatric disorders. Exome studies noting modest enrichment of rare inherited variants as well. The relative enrichment of de novo PTVs in cases varies by ascertainment strategy: burden is greatest in individuals with developmental delay (DD), intellectual disability, or multi-system congenital anomalies; moderate in individuals with ASD or isolated developmental anomalies; and lowest in schizophrenia and other neuropsychiatric disorders. Indeed, hundreds of genes have now been discovered across this spectrum of developmental disorders, with associations driven by phenotype severity and cohort size.

Furthermore, it has been well established that ASD and DD cases harbor an excess of very large copy number variants (CNVs) compared to unaffected siblings. While these CNVs of large genomic segments represent an approximately 3.5-fold increase in ASD risk, their incorporation into genomic studies has long represented a significant technical challenge. Early studies using microarrays were limited to CNVs of hundreds of kilobases to megabases. Among the most significant findings from these studies was the recurrent genomic disorder (GD) loci associated with syndromic phenotypes that arose due to mispairing of long homologous segments, a mechanism known as non-allelic homologous recombination (NAHR). Due to their high mutation rate, these reciprocal GD segments, a mechanism known as non-allelic homologous recombination (NAHR), are depleted of functional variation in the broader population. Due to their high mutation rate, these reciprocal GD segments, a mechanism known as non-allelic homologous recombination (NAHR), are depleted of functional variation in the broader population. What remains less clear is the relative burden of CNVs at the resolution of individual genes or exons in ASD etiology, as studies to date have been limited by the relatively small samples assessed by whole-genome sequencing (WGS) and the high false positive rates for existing exome-based CNV discovery methods. Prior studies by the Exome Aggregation Consortium (ExAC), 100 Genomes Project, and Genome Aggregation Database (gnomAD) have established size distributions and mutation rates for structural variants (SVs) across human populations, while studies leveraging long-read WGS have provided further insights into the broader swath of human SVs, over 99% of which are not readily detectable by microarray. Here, we demonstrate that significant improvements in computational models to normalize exome sequencing read depth, combined with appropriate quality control filters, enable discovery of even small rare coding CNVs with validation rates comparable to indels. These methods allow us to build upon prior studies to now jointly analyze rare coding SNVs, indels, and CNVs at the resolution of individual genes and exons from large-scale ASD datasets.

Rare variant discovery can also be enhanced through the integration of functional data, particularly in cohorts such as ASD that display strong selective pressure and reduced fecundity. One such measure is the recently developed loss-of-function observed/expected upper bound fraction (LOEF) score, which is a continuous measure of selective pressure against PTVs in each gene. Similarly, the ‘missense badness’, PolyPhen-2, and constraint (MPC) score is one of several measures of the estimated deleteriousness of missense variation. In this study, we used a Bayesian statistical framework, the Transmission and De Novo Association (TADA) model, to incorporate these functional annotations and jointly analyze 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). Our analyses 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 cell types in which they are enriched during early neuronal development. 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 coding variants in a large ASD exome dataset

We aggregated whole exome sequencing (WES) data across 33 ASD research cohorts, totaling 63,237 individuals: 15,036 affected offspring, 5,492 unaffected offspring, 28,522 parents, as well as 5,591 affected and 8,597 unaffected individuals from case-control studies (one sample exists in our dataset as both mother and affected offspring). Fig. 1a, Supplementary Tables 1, 2, 3, 4.1). Data from 35% of these individuals had not previously been analyzed or published. All accessible DNA sequence reads were aligned to the GRCh38 human reference genome and coding SNVs, indels, and CNVs were identified using GATK haplotype caller 42 while a new method, GATK-gCNV, was used for CNV delineation. Variant counts were consistent across cohorts, with an average of 1.64 (1.66/affected, 1.57/unaffected) de novo SNVs, 0.18 (0.18/affected, 0.16/unaffected) de novo indels, and 0.035 (0.042/affected, 0.014/unaffected) de novo CNVs per offspring.

Consistent with prior studies, we observed more PTVs and damaging missense variants in individuals with ASD compared to unaffected individuals (Fig. 1b-c). Enrichment of de novo and inherited PTVs in ASD cases was greatest in genes with the strongest evidence of selective constraint, represented by low LOEUF scores39 (Supplementary Tables 4.2-4.7). Both de novo and inherited PTVs were enriched in the top three deciles of LOEUF (binomial test; 5,446 total genes; Fig. 1b). We grouped missense variants into three bins based on MPC score, which we termed MisB (MPC ≥ 2) and MisA (2 > MPC ≥ 1), with MPC < 1 for all remaining missense variants. MisA variants and, to a lesser degree, MisB variants were significantly enriched in ASD individuals (Fig. 1c). Our aggregated cohort demonstrated the greatest risk from de novo variation, with modest but significant ASD risk observed in rare case/control (for which de novo status cannot be determined) and inherited PTVs in the top three deciles of constrained genes, as well as in MisB and MisA variants.

Detection and analysis of rare and de novo CNVs in ASD

Previous studies using chromosomal microarrays have established a clear etiological role for large, rare and de novo CNVs in ASD cases compared to their unaffected siblings and the general population21,22,25–27,44–46. Despite their considerable impact on ASD risk, CNV discovery at the resolution of individual exons from exome sequencing has represented a significant technical challenge. To overcome these challenges, we applied and benchmarked a pipeline built around GATK-gCNV, a novel Bayesian

Fig. 1 | Overview of SNV/indel and CNV rates in ASD by mode of inheritance.

a, The ASD cohort for SNV/indel analysis consisted of 49,049 family-based samples (15,036 cases) and 14,188 case/control samples (5,591 cases). 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, we found significant enrichment in cases for PTVs among the most constrained genes (lower LOEUF deciles), which weakened as negative selection against PTVs is relaxed (higher LOEUF deciles). c, Equivalent information and statistical significance for missense and synonymous variants. Synonymous variants were not enriched in cases or controls, as evaluated via binomial tests, in keeping with well-calibrated variant calling. d, The GATK-gCNV exome CNV detection pipeline when compared against WGS on overlapping samples achieved a sensitivity of 82% and positive predictive value of 89% for rare CNVs (<1% site frequency) that spanned ≥ 4 more callable exons (red line). e, The relative difference in variant frequency between cases and controls for deletions. Using binomial tests, we found that deletions overlapping genes in the lowest LOEUF decile were observed at an elevated rate in affected individuals. f, Equivalent analysis for duplications, demonstrating a similar pattern of enrichment compared to deletions but with a weaker effect.

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; Del: Deletion CNV; Dup: Duplication CNV. Statistical tests: b,c,e,f: binomial test.
read depth-based CNV discovery tool for short-read sequencing data. Our pipeline includes a PCA-based method for inferring large systematic differences, such as different exome capture kits (Supplementary Fig. 1), and we performed extensive benchmarking of this method from 7,165 samples from ASD quartet families for which matching gold-standard CNV calls from microarray and WGS were available. We measured sensitivity and positive predictive value (PPV) as a function of the number of captured exons from canonical protein-coding transcripts. Our application of GATK-gCNV achieved a sensitivity of 82% and a PPV of 89% for rare CNVs (site frequency < 1%) that spanned 3 or more such exons (Fig. 1d, Supplementary Fig. 3). When we considered de novo CNVs detected by WGS, GATK-gCNV achieved 82% sensitivity and 97.5% PPV for de novo CNV discovery at three-exon or greater resolution (Supplementary Fig. 3).

We applied GATK-gCNV to the 57,294 samples from the ASD cohort for which raw sequencing data were available (Supplementary Methods, Supplementary Table 3), including 13,694 affected and 5,007 unaffected offspring with data from both parents and 608 cases and 11,312 controls from non-family based cohorts. We focused our analysis on high quality CNVs spanning at least three captured protein-coding exons (Supplementary Methods), resulting in 16,889 rare (frequency < 1%) and 650 de novo autosomal CNVs, corresponding to an average of 0.90 (0.91/affected, 0.89/unaffected) rare and 0.035 (0.042/affected, 0.014/unaffected) de novo CNVs per offspring.

Consistent with de novo coding CNVs mediating substantial risk for ASD, we observed that 3.93% of ASD cases and 1.40% of unaffected siblings harbored at least one de novo coding CNV, respectively (odds ratio [OR]: 2.88, p = 3.9 x 10-4, Fisher’s exact). Splitting the CNVs by dosage (deletions versus duplications) and assigning each event to a LOEUF decile based on the most constrained gene within its boundaries, we observed that de novo deletions with at least one PTV intolerant gene showed the greatest degree of enrichment in ASD cases across all variant classes evaluated (OR: 9.10, p = 2.81 x 10-20, Fisher’s exact, Fig. 2c). Enrichment of deletions was also observed in case/ control but not in rare inherited analyses. A similar but less significant pattern was observed for duplications across constraint deciles and modes of inheritance (Fig. 1e-f). Additionally, with breakpoints restricted within a gene have been suggested to harbor risk differently than duplications of an entire gene, likely through a loss of function mechanism. Leveraging our de novo duplication data, we found that partially duplicated genes in cases harbored more de novo PTV evidence compared to partially duplicated genes in controls (1.3 fold, p = 1.7 x 10-3, binomial test).

To characterize the impact of exon and gene-level resolution in CNV discovery, we first considered 75 large genomic segments (Supplementary Table 4.8) associated with NDDs from curated literature. Of the 650 de novo CNVs discovered from WES here, 223 (34.3%) matched one of these loci with at least 50% reciprocal overlap, including 109 de novo GD deletions in cases compared to 6 in controls (OR: 6.7, p = 3.9 x 10-4, Fisher’s exact) and 100 de novo GD duplications in cases compared to 8 in controls (OR: 4.6, p = 5.4 x 10-3, Fisher’s exact, Fig. 2b). Considering rare inherited CNVs, we observed a non-significant trend towards enrichment of GD CNVs in cases (OR: 1.3, p = 0.07, Fisher’s exact). Excluding the GD loci, the remaining 427 de novo CNVs were also enriched in cases but with more modest effect sizes (de novo non-GD deletion: OR = 2.8, p = 1.5 x 10-3, Fisher’s exact; de novo non-GD duplications: OR = 1.8, p = 5.1 x 10-3, Fisher’s exact). A greater degree of enrichment was observed for non-GD deletions that overlapped a constrained gene (OR: 5.10, p = 9.7 x 10-11, Fisher’s exact, Fig. 2c) than for those that did not (OR: 1.32, p = 0.32, Fisher’s exact, Fig. 2c) or for de novo PTVs in constrained genes (OR: 2.74, p = 8.5 x 10-34, Fisher’s exact, Fig. 2c). Effect sizes for de novo duplications were consistently smaller, although similar patterns were observed in relation to GD loci and constrained genes (Fig. 2c). Finally, when we considered de novo GD CNVs that altered one of 71 genes associated with ASD risk (FDR < 0.015 non-missing), we detected observed 48 variants in cases (29 deletions, 17 duplications) compared to none in siblings (p = 8.5 x 10-17, Fisher’s exact). Considering the list of 373 genes implicated in both ASD and DD risk (see below) this increased to 130 de novo CNVs in cases and one in siblings (OR: 49.0, p = 1.1 x 10-16, Fisher’s exact).
Consistent with a female protective effect in ASD, which proposes that a higher burden of risk factors is required for an ASD diagnosis in females\textsuperscript{17,55}, the burden of \textit{de novo} CNVs overall was higher in female cases than male cases (5.94\% vs. 3.47\%, OR:1.76, p=3.4x10\textsuperscript{-8}, Fisher’s exact). This effect was more pronounced for CNVs than PTVs (p=2.6x10\textsuperscript{-5}, logistic regression) or missense variants (p=2.0x10\textsuperscript{-2}, logistic regression, Fig. 2d). While \textit{de novo} SNVs and indels frequently arise on the paternal allele\textsuperscript{44,46}, a maternal allele bias has been observed from \textit{de novo} CNVs in ASD\textsuperscript{47}. Using parental SNP data, we estimated the allelic parent-of-origin for 330 \textit{de novo} CNVs (\textit{Supplementary Methods}) and observed no overall bias (Fig. 2e, 130 maternal vs. 144 paternal, p=0.43, binomial test). However, restricting analyses to \textit{de novo} CNVs at NAHR-mediated GD loci, we recapitulated prior findings with 70\% arising on the maternal allele (70 maternal vs. 30 paternal, p = 7.9x10\textsuperscript{-5}, binomial test). Including 20 additional samples from the Simons Searchlight project\textsuperscript{48}, we find that 94.3\% of 16p11.2 CNVs occur on the maternal allele (Fig. 2e, 33 maternal vs. 2 paternal, p=3.7x10\textsuperscript{-6}, binomial test). In contrast, CNVs at non-NAHR-mediated GD loci showed a 65.5\% paternal bias (Fig. 2e, 60 maternal vs. 114 paternal, p=5.2x10\textsuperscript{-5}, binomial test), consistent with a mechanistic bias in CNV formation and previous findings of a paternal origin across all non-recurrent/non-pathogenic CNVs\textsuperscript{49}.

\textbf{Integrated discovery of ASD-associated genes across variant types and inheritance classes}

Our analyses identified enrichment of rare protein-coding variants in ASD cases with effect sizes varying by mode of inheritance, variant class (PTV, misB, misA, deletion, duplication), and evolutionary constraint. We sought to utilise these insights and the size of our cohort (63,237 samples) to refine gene discovery in ASD by extending the TADA analytic framework\textsuperscript{10,22,41} to include: (1) rare and \textit{de novo} CNVs, (2) variants present in unaffected offspring, and (3) a continuous measure of evolutionary constraint from gnomAD (LOEUF\textsuperscript{39}, \textit{Supplementary Methods, Supplementary Fig. 4}). A Bayes Factor (BF) was calculated to represent evidence of association for each autosomal protein-coding gene across variant types and modes of inheritance, taking into account null mutation rates and prior relative risks, when appropriate (Fig. 3).

Applying this model to the aggregated ASD data (TADA-ASD) we identify 71 genes associated with ASD at a FDR ≤ 0.001 and 183 genes associated at FDR ≤ 0.05 (\textit{Supplementary Table 4.9}). Notably, the FDR ≤ 0.001 threshold is approximately equivalent to Bonferroni significance when back-calculating a p-value and correcting for 16,128 autosomal genes (p < 2.8x10\textsuperscript{-8}), making it comparable to other recent studies of schizophrenia\textsuperscript{13} and DD\textsuperscript{11}. We focus on these 71 genes here, and illustrate in Fig. 3 that 93\% of these genes have contributions from multiple variant classes (Fig. 3b). They also represent a meaningful increase in ASD discovery - a prior study from Satterstrom et al. included a subset of 11,986 of the cases aggregated here, producing 32 ASD risk genes at this same FDR ≤ 0.001 threshold using the TADA parameters applied in that study, which increases to 51 genes when incorporating the updated model parameters and the CNVs discovered here (Fig. 3c). The BFs within TADA allow us to assess genomic architecture through the relative contributions across variant types

\textbf{Integrating variant types and inheritance classes significantly boosts association power and reveals mutational biases}

\textbf{a.} Our implementation of the TADA model includes \textit{de novo}, case/control, and rare inherited modules for each variant type: PTV, MisB, MisA, deletion, and duplication. We leverage information from ASD probands as well as unaffected siblings in evaluating the effect of \textit{de novo} variants. \textbf{b.} The evidence of ASD association contributed by each variant type for each of the 71 ASD genes with FDR ≤ 0.001. Some genes are predominantly associated through missense variants and duplications (e.g., PTEN, SLC6A1), suggesting mechanisms such as gain-of-function in contrast to haploinsufficient loss-of-function for genes with large contributions from PTVs and deletions. \textbf{c.} Applying TADA to our aggregated ASD dataset yields 71 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 makes more efficient use of the available evidence of association by integrating more information into our statistical modeling. \textbf{d.} Across 71 ASD genes, the majority of evidence for ASD association is derived from PTVs and MisB variants and, e, \textit{de novo} variants.

\textbf{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: a Extended TADA model.}

---

**Fig. 3** Integrating variant types and inheritance classes significantly boosts association power and reveals mutational biases

\textbf{a.} Our implementation of the TADA model includes \textit{de novo}, case/control, and rare inherited modules for each variant type: PTV, MisB, MisA, deletion, and duplication. We leverage information from ASD probands as well as unaffected siblings in evaluating the effect of \textit{de novo} variants. \textbf{b.} The evidence of ASD association contributed by each variant type for each of the 71 ASD genes with FDR ≤ 0.001. Some genes are predominantly associated through missense variants and duplications (e.g., PTEN, SLC6A1), suggesting mechanisms such as gain-of-function in contrast to haploinsufficient loss-of-function for genes with large contributions from PTVs and deletions. \textbf{c.} Applying TADA to our aggregated ASD dataset yields 71 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 makes more efficient use of the available evidence of association by integrating more information into our statistical modeling. \textbf{d.} Across 71 ASD genes, the majority of evidence for ASD association is derived from PTVs and MisB variants and, e, \textit{de novo} variants.

\textbf{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: a Extended TADA model.}
and modes of inheritance (Fig. 3d-e, Fig. 4a-b). In keeping with haploinsufficiency as the primary pathogenic mechanism, PTVs and deletions account for over 90% of the evidence in 20 of the 71 ASD-associated genes (28.2%). However, for 10 genes (14.1%), over 90% of the evidence comes from missense variants and duplications (e.g., DEAF1, SLC6A1; Fig. 4a). Included within these 10 genes is PLXNA1, for which the evidence comes primarily from inherited missense variation localized within the Plexin domain of the encoded protein (Fig. 4b-c).

We next sought to estimate the magnitude of the ASD risk imparted by the genes discovered here and described in Fig. 3. Within the 71 genes, de novo PTVs or damaging missense variants were detected in 4.0% of cases and 0.5% of controls, giving a combined OR of 8.19. Using cross validation (Supplementary methods), we refined this estimate by variant type and mode of inheritance with OR estimates ranging from 16.0 for de novo PTVs to 1.02 for inherited MisA (Supplementary Table 4.10). We also applied this cross validation approach to estimate the yield of future gene discovery efforts, predicting a greater than linear increase in gene discovery at this same statistical threshold from a two-fold increase in the current sample size (150 genes; 95% CI: 108-278) or a three-fold increase (344 genes; 95% CI: 218-500).

**Comparing the genetic architectures of ASD and general NDDs**

ASD is frequently comorbid with other syndromic and non-syndromic NDDs, including intellectual disability (ID), and neurological disorders, including seizures⁴⁶–⁵⁰. Prior analyses have clearly demonstrated both overlapping and distinct genetic architectures across ASD and DD, with significant heterogeneity across these broadly defined NDDs. To explore commonalities in genes that impact such NDDs, we integrated our ASD data with an independent cohort of 91,605 family-based samples encompassing 31,058 offspring with DD, the vast majority of which (>95%) were diagnosed with at least one NDD¹¹. Exome sequencing from this DD cohort was recently analyzed by the Deciphering Developmental Disorders (DDD) project using DeNovoWest, a permutation-based frequentist method, and reported exome-wide significant associations between rare SNVs and indels in 252 autosomal genes. However, these analyses did not include case-control samples, rare inherited SNVs and indels, or CNVs. We sought to re-analyze these DDD data to enable direct comparisons between the ASD and DD cohorts using uniform statistical models and significance thresholds between studies.

We applied the extended TADA framework to de novo variants in the DD cohort (TADA-DD) and found 309 autosomal genes associated at an FDR threshold of ≤ 0.001, including 94% of the initial 252 autosomal genes in Kaplanis et al. 2020 (Supplementary Table 4.9, TADA-DD found 477 at FDR < 0.05). We observed the FDR values generated by our implementation of TADA (TADA-DD) to be highly correlated to those derived from the DeNovoWest significance values in Kaplanis et al. 2020¹¹ (r=0.95, p<10x10^{-22}, Supplementary Fig. 5). Given the enrichment of cases with severe and syndromic disorders in this DD cohort compared to the ASD cohort⁵⁰–⁵¹, the

---

**Fig. 4 | Relative contribution of evidence types in ASD risk genes**

a. For 71 ASD-associated genes (FDR ≤ 0.001) the relative evidence of ASD association in the extended TADA model (log10BF) is shown for variants acting via a likely loss-of-function (LOF) 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. Equivalent plot of the relative evidence from de novo (x-axis) versus inherited (y-axis) variation for the same 71 ASD-associated genes. c. Evidence for ASD-association for the gene PLXNA1 is derived from de novo and transmitted missense variants, especially in the Plexin domain at the C-terminus of the Plexin-A1 protein. Abbreviations: BF: Bayes factor. Statistical tests: c: Transmission disequilibrium test.

---

Since the first unbiased whole-genome analyses in ASD⁴², a notable overlap has been observed between genes affecting ASD and those affecting development more broadly, including intellectual development. Thus, it is reasonable to conjecture that integrating evidence for association from ASD and DD cohorts could yield additional genes influencing risk for ASD and broader development. Still, a cardinal rule of such meta-analysis is that the data should not be too heterogeneous. To determine whether the genes identified in the ASD cohort were also associated in the DD cohort, and vice versa, we converted the distribution of TADA q-values to p-values for each study. If the genes identified as associated in the DD cohort were also associated with ASD, the distribution of their ASD association p-values would be skewed toward zero; if they were not associated, the p-values would follow a uniform distribution; and, if some were associated and others were not, the p-values would follow a mixture distribution. Selecting the genes associated in the DD cohort at FDR ≤ 0.05, we evaluated the distribution of their p-values in the ASD cohort. The estimated proportion of associated genes (i.e., mixing parameter) was 0.701 (limma::propTrueNull, Fig. 5c), indicating that 70.1% of these genes affect risk for ASD. The converse analysis conditioning on the ASD-associated genes suggests that 86.3% of ASD risk genes have some broad effects on development (limma::propTrueNull, Fig. 5d).
evidence for association in the DD cohort at the same threshold (Fig. 5d). Moreover, while the remaining 86.3% of the genes are likely pleiotropic\(^6\), some could have a far greater impact on risk for ASD while others have greater impact on risk for other features of development. To evaluate the overall gene-set heterogeneity between the ASD and DD cohorts, we retained only \textit{de novo} SNVs/indels for independent gene-level BF calculations. For the 373 genes at TADA-ASD FDR ≤ 0.001, we observed a Pearson’s correlation of 0.78 (p=2.4x10\(^{-17}\)) between the two major subcomponents of the ASD cohort (ASC+SSC versus SPARK), compared to only 0.42 (p=2.4x10\(^{-17}\)) between the BFs of the ASD and DD cohorts. These observations signal much more consistent gene-level evidence between the ASD subcohorts than between the ASD and DD cohorts (Supplementary Fig. 6) and reflect both differential and shared genetic architecture underlying these often-comorbid phenotypes.

Next, we determined if certain genes were observed to be more commonly mutated in one cohort or the other. For these analyses, we first identified 464 “signal genes,” defined as any gene with FDR ≤ 0.05 in either TADA-ASD or TADA-DD from \textit{de novo} PTV and MissB variant data (which showed the greatest association and were least likely to be benign). Of the 27 signal genes, 120 belonged to TADA-ASD, 428 to TADA-DD, and 84 to both. Notably, even for the 84 genes significant in both cohorts, we also observed significant variant count heterogeneity (X2=317.6, DF=83, p=3.8x10\(^{-23}\)) between the cohorts.

Given this heterogeneity in mutational patterns, we sought to capture which of the 464 genes have more variation in either the ASD or the DD cohorts. A common way to assess this is by a standardized chi-squared test statistic (C statistic), but its power to discriminate is abrogated by the much higher burden of risk variants in the DD than ASD samples (Fig. 5a-b). We therefore adjusted for the difference in mutational burden between the cohorts by downsampling the DD mutations to be equal to the cohort of ASD mutations, then computing the C statistic. Genes with fewer than expected mutations in the ASD cohort relative to the DD cohort achieved a negative C statistic; a positive C statistic indicated overrepresentation in the ASD cohort. Next, a mixture model was adopted to disentangle the two commingled distributions, which assigns posterior probabilities that a gene is from the ASD or DD component of the statistical distribution (Fig. 5c-f). Using a posterior probability cutoff of greater than 0.99, we find 36 genes to be a part of the ASD mixture component and 82 genes to be a part of the DD component (Fig. 5f).

\textbf{Mutational enrichment identifies differential neuronal layers impacted by ASD and DD genes}

To gain insight into differences in expression of genes identified from the ASD versus DD cohorts, we examined single-cell gene expression patterns from human fetal brains. Two studies provided more than 37,000 cortical cells ranging from 6-27 weeks post-conception\(^{56,57}\). To integrate these data, we first removed strong batch effects due to different protocols using cFIT and then obtained integrated factor loadings and gene expression for all measured cells\(^{56}\). UMAP plots visualizing all cells from two sources, after removal of batch effects, showed that similar cell types from the different batches group together, while cells unique to either batch were also preserved\(^{57}\) (Fig. 6a, Supplementary Fig. 7).

With the integrated data, we applied unsupervised clustering to identify cell subtypes in the context of a hierarchical tree to illustrate the relationships of major and minor cell type clusters. Using the MRtree method\(^{67}\), we observed that cells of each labeled type were merged across both 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. Likewise, minor splits reflected the expected relationships between cell types (Supplementary Fig. 8a). Based on the trajectory analysis of Poliadakis et al.\(^{58}\), the ExM clade is less mature than the ExM clade, which in turn is less mature than the ExMU clade.

\textbf{Heterogeneity of mutation patterns between ASD and DD associated genes}

Due to the frequent comorbidity of ASD and DD phenotypes, isolating genes that exert a greater effect on ASD than they do on other DDs has remained challenging. Still, as documented above, 13.7% of the TADA-ASD FDR ≤ 0.05 genes show little
We subsequently assessed any differentials in strength of enrichment of the DD- and ASD-predominant risk genes within cell clusters for genes meeting the posterior probability 0.99 threshold (Fig. 5f). Of the 36 genes classified as ASD-predominant, 18 were expressed in these cell types; and, of the 82 classified as DD-predominant genes, 51 were expressed. Using odds ratio to reflect the strength of signal, we observed that both DD- and ASD-predominant genes were enriched in both interneurons and excitatory neurons compared to glial and progenitor cells (Fig. 6b, Supplementary Fig. 9, Supplementary Tables 4, 12-4, 13). While the enrichment was broadly distributed, the signal for DD-predominant enrichment tended to be found for cell types appearing early in the lineages of neurons. DD-predominant genes highlight four cell types, namely intermediate progenitors (IP), precursors of interneurons from the caudal ganglionic eminence (InCGE), migrating excitatory neurons (ExN3), and maturing excitatory neurons (ExM2). By contrast, ASD-predominant genes are strongly enriched in only one cell type, maturing excitatory neurons (ExMU1) and its clade. In keeping with a more mature neuron type, the ASD-predominant genes play roles in cytoskeleton organization, synaptic signaling, and activity-dependent programs. Our results regarding ASD agree with the conclusion of Polioudakis et al., which was based on a subset of these data. If we judge enrichment solely by significance after Bonferroni correction for 21 cell types, these conclusions still hold: 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 mature cells than ASD-predominant genes.

**Emergence of shared ASD schizophrenia risk genes**

The genetic risk for ASD has also been suggested to overlap with that of other neuropsychiatric disorders, especially schizophrenia, such that the joint study of ASD and schizophrenia might reveal additional insights into both disorders. The Schizophrenia Exome Meta-Analysis (SCHEMA) Consortium recently analyzed exome sequencing from 24,248 schizophrenia cases and 97,322 controls, identifying 10 genes where ultra-rare coding variants were associated with schizophrenia after Bonferroni correction, and 244 genes at $p < 0.01$. We compared our ASD- and NDD-associated genes to the SCHEMA findings 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. Among the 71 ASD genes we discovered at an FDR ≤ 0.001, we found that 63 show an association with DD (using FDR ≤ 0.05, based on TADA-DD), and 8 show an association with schizophrenia (using $p < 0.01$, based on values reported by Singh et al.). If the two associations were independent, we would expect $\frac{8 \times 71}{2} = 284$ of the 8 ASD schizophrenia genes to also show an association with DD (based on 63/71 = ~88% of the ASD genes overlapping DD). However, we find 4 of the ASD schizophrenia genes (NRXN1, ANK2, BRSK2, and DSCAM) lack an association with DD, which is a significant overrepresentation compared to random chance alone ($p = 0.006$, binomial test). This suggests that one subset of ASD risk genes may overlap DD, while a different subset overlaps schizophrenia.

We also analyzed these findings with an alternative approach: the distribution of genes in Fig. 5e (the 464 genes with FDR ≤ 0.05 for ASD and/or DD in our heterogeneity analysis) gives 36 ASD-enriched genes and 82 DD-enriched genes. Of the 244 genes identified by SCHEMA as schizophrenia-associated at $p < 0.01$, 6 genes (ANK2, ASH1L, BRSK2, CGREF1, DSCAM, and NRXN1) overlap the 36 ASD genes, while only 3 (ATP2B1, GRIN2A, and HIST1H1E) overlap the 82 DD genes. If we compare to the null hypothesis that each of the genes in our TADA model has an equal chance of being schizophrenia-associated, then the ASD-schizophrenia overlap is significantly enriched ($p = 8.5 \times 10^{-6}$, binomial test), while the DD-schizophrenia overlap is not ($p = 0.10$, binomial test). The two outcomes (6/36 vs. 3/82) are also significantly different when compared to each other ($p = 0.023$, Fisher’s exact test), reinforcing the idea that the shared genetic risk between ASD and schizophrenia may be distinct to that observed between ASD and DD.

**Discussion**

We report here a gene discovery effort in ASD that integrates coding SNVs, indels, and exome-based CNVs in ASD. These analyses revealed an allelic spectrum associated with ASD that was dominated by de novo PTVs, damaging missense variants, and deletions within genes that display evidence of intolerance to loss-of-function variation in the general population. Nonetheless, of the 71 genes identified as associated at FDR ≤ 0.001, all included association evidence from multiple inheritance classes and 10 displayed the strongest evidence from de novo missense variants and duplications, including one gene (PLXNA1) with domain-specific transmission distortion. The proportion of log BF evidence arising from PTVs and DELs steadily decreased from 69.2% among the 71 TADA-ASD risk genes to 51.1% and 41.0% for FDR ranges of (0.001, 0.01) and (0.01, 0.05).
respectively, while inherited evidence increased from 4.5% to 10.1% and to 20% for the same FDR ranges, indicating that future discoveries are likely to arise from the more subtle effects associated with missense and inherited variation as sample sizes continue to increase.

This study also presents the largest exploration to date of individual gene and multi-exon resolution CNVs from exome sequencing and their contributions to ASD architecture. In a subset of the SSC quartets in this study, we found that >80% of all CNVs discovered by WGS and spanning three or more exons could be recapTUREd by exome-based CNV detection, while >90% of all predicted de novo CNVs from exome sequencing at this resolution were confirmed by WGS and/or molecular validation. We readily recapitulated the observed excess of CNVs in known GD regions characterized from microarray studies, and quantified the relative effect size of CNVs outside of GD regions in comparison to PTVs. Prior studies have reported an enrichment of CNVs of maternal origin in ASD probands15. Here, we found a strong bias in the gamete-of-origin giving rise to de novo CNVs based on the mechanism of formation. Most NAHR-mediated CNVs across all GD arose on the maternal chromosome (e.g., 95% of 16p11.2 CNVs), whereas the majority of all CNVs, which are driven by non-homologous or microhomology mediated repair, were preferentially inherited in ASD.”

While both are highly associated with ASD, the combined impact of rare and common variant polygenic risk across males and females, and combined gene discovery from ASD and DD datasets15,17,18. As sample sizes continue to grow rapidly, we expect that the framework presented here will continue to yield returns in both gene discovery and improved understanding of the differential risks to the disorders on the neurodevelopmental and neuropsychiatric spectrum posed by variants within the same genes.

**Acknowledgements**

We thank all of the individuals who participated in this research. We also thank all contributors 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) project, 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., #575326 to M.E.T. and M.J.D., and #571009 to J.D.), the SPARK project and SPARK analysis projects #606362 and #6008540 to M.E.T., M.J.D., J.B., B.D., K.R., S.J.S., SFAIR #042281 to S.J.S., M.W.S., M.K.R., and #647371 to S.J.S.), NHGRI (HG008895 to M.J.D., S.G., M.E.T.), NIMH (MH151957 to M.E.T., MH111658 and MH057881 to B.D., MH111661 and MH100233 to J.D.B., MO1 MH09990 to K.R., MH111660 to M.J.D., and MH111662 and MH100027 to S.J.S. and M.W.S.), NICHD (HD061256 to M.E.T.), AMED (JP21WM042507 to N.O), and the Seaver Foundation (J.B. was supported by an Autism Speaks Postdoctoral Fellowship and R.L.C. was supported by NSF GRFP #2017240332).


