npj | Genomic Medicine

ARTICLE OPEN

Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes

Pamela Feliciano[1], Xueya Zhou[2], Irina Astrovskaya[1], Tychele N. Turner[3], Tianyun Wang[3], Leo Brueggeman[4], Rebecca Barnard[5], Alexander Hsieh[2], LeeAnne Green Snyder[1], Donna M. Muzny[6], Aniko Sabo[6], The SPARK Consortium, Richard A. Gibbs[6], Evan E. Eichler[3,7], Brian J. O'Roak[5], Jacob J. Michaelson[4], Natalia Volfovsky[1], Yufeng Shen[2] and Wendy K. Chung[1,8]

Autism spectrum disorder (ASD) is a genetically heterogeneous condition, caused by a combination of rare de novo and inherited variants as well as common variants in at least several hundred genes. However, significantly larger sample sizes are needed to identify the complete set of genetic risk factors. We conducted a pilot study for SPARK (SPARKForAutism.org) of 457 families with ASD, all consented online. Whole exome sequencing (WES) and genotyping data were generated for each family using DNA from saliva. We identified variants in genes and loci that are clinically recognized causes or significant contributors to ASD in 10.4% of families without previous genetic findings. In addition, we identified variants that are possibly associated with ASD in an additional 3.4% of families. A meta-analysis using the TADA framework at a false discovery rate (FDR) of 0.1 provides statistical support for 26 ASD risk genes. While most of these genes are already known ASD risk genes, *BRSK2* has the strongest statistical support and reaches genome-wide significance as a risk gene for ASD (p-value = 2.3e−06). Future studies leveraging the thousands of individuals with ASD who have enrolled in SPARK are likely to further clarify the genetic risk factors associated with ASD as well as allow accelerate ASD research that incorporates genetic etiology.

npj Genomic Medicine (2019)4:19 ; https://doi.org/10.1038/s41525-019-0093-8

INTRODUCTION

Autism spectrum disorder (ASD) is an extremely variable condition characterized by deficits in social interactions and restrictive, repetitive behaviors. Currently, there are no FDA approved medications that address these core symptoms, despite the life-long morbidity and increased mortality in adults with ASD.[1]

Despite the significant clinical heterogeneity of this condition, many studies have shown that ASD is highly heritable, with genetic risk factors thought to explain the majority of the risk for ASD.[2] Over the past decade, genomic studies focused on de novo, likely gene disrupting (dnLGD) variants (stopgain, frameshift, and essential splice site) have identified ~100 high-confidence ASD risk genes or loci.[3,4] Previous studies have identified molecular diagnoses in 6–37% of individuals with ASD, with higher yields in individuals with additional co-morbidities that include intellectual disabilities, seizures, and other medical features.[5]

Here we describe the results of a pilot study that genetically characterized 457 families with one or more members affected with ASD enrolled online in SPARK.[6] SPARK's mission is to create the largest recontactable research cohort of at least 50,000 families affected with ASD in the United States for longitudinal phenotypic and genomic characterization who are available to participate in research studies. Using exome sequencing and genome-wide single nucleotide polymorphism (SNP) genotyping arrays, we identified variants that are the likely primary genetic cause of ASD in 14% of families. We also demonstrated that the genetic architecture in this self-reported cohort is similar to published, clinically confirmed ASD cohorts.[3,4,7] Combining the SPARK data with prior studies, our analyses provide strong evidence that *BRSK2* is a high-confidence ASD risk gene (FDR q-value = 0.0015) and provide evidence that strengthens the association of additional genes (*FEZF2, ITSN1, PAX5, DMWD*, and *CPZ*) in ASD.

RESULTS

Variant discovery

We report the exome sequencing and genotyping results of 1379 individuals in 457 families with at least one offspring affected with ASD, including 418 simplex and 39 multiplex families (Supplementary Fig. 1). Over 80% of participants are predicted to have European ancestry based on principal component analysis of common SNP genotypes (Supplementary Fig. 2). The male to female ratio is 4.4:1 among 418 offspring cases in simplex families, and 2.9:1 among 47 offspring cases in multiplex families. Of the 465 offspring affected with ASD, 25.6% also reported intellectual disability (Table 1). We identified 647 rare (allele frequency (AF) <0.001 in ExAC v0.3) de novo single nucleotide variants (SNVs) and

[1]Simons Foundation, New York, NY 10010, USA; [2]Department of Systems Biology, Columbia University, New York, NY 10032, USA; [3]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA; [4]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, IA 52242, USA; [5]Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97239, USA; [6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA; [7]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA and [8]Department of Pediatrics, Columbia University Medical Center, New York, NY 10032, USA
Correspondence: Pamela Feliciano (pfeliciano@simonsfoundation.org) or Wendy K. Chung (wkc15@cumc.columbia.edu)

A full list of The SPARK Consortium members appears at the end of this paper.

These authors contributed equally: Pamela Feliciano, Xueya Zhou, Irina Astrovskaya, Tychele N. Turner

Received: 20 December 2018 Accepted: 11 July 2019
Published online: 23 August 2019

indels (Supplementary Data 1) in coding regions and splice sites (1.4/offspring), including 85 likely gene disrupting (LGD) variants and 390 missense variants. Similar to the de novo variants identified from 4773 clinically ascertained ASD trios from previous studies,[3–8] the frequency of dnLGD variants in the 465 affected offspring in SPARK (0.18/offspring) is 1.76-fold higher than the baseline expectation calculated by a previously published mutation rate model[9] (p-value = 1.2×10^{-6} by one-sided exact Poisson test) ("Methods"; Supplementary Data 2).

To identify de novo missense variants that are likely damaging, we applied two deleterious missense (D-mis) prediction algorithms on published ASD and SPARK de novo variants. Among the 390 de novo missense variants in affected offspring, 43.6% are predicted to be deleterious using CADD score ≥25[10] and show 1.28-fold enrichment compared with baseline expectation in the general population (p-value = 6.6×10^{-4} by one-sided exact Poisson test). Using a more strict D-mis prediction algorithm with MPC score ≥2,[11] 8% of de novo missense variants are predicted as deleterious and are enriched 1.88-fold in affected offspring which is comparable with the enrichment of dnLGD variants (p-value = 9.9×10^{-4} by one-sided exact Poisson test). The overall burden of de novo D-mis variants is similar to published studies (Supplementary Data 2).

Variants in constrained genes (pLI ≥ 0.5)[12] explain most of the burden of dnLGD variants and de novo D-mis variants (defined by an MPC score ≥2) in the affected offspring in our study (Supplementary Data 2,b). Consistent with previous findings supporting the female protective model,[13] we observed a nonsignificant trend toward a higher frequency of dnLGD variants in constrained genes in female cases compared with males (0.135/female vs 0.096/male), as well as higher frequency of de novo D-mis variants in female cases (CADD ≥25: 0.416/female vs 0.354/male, MPC ≥2: 0.09/female vs 0.066/male).

We also investigated deleterious inherited SNV/indel variants and found a modest excess of transmitted, rare LGD (AF < 0.001 in ExAC v0.3) variants observed only once among parents in our cohort (singletons) in constrained genes with pLI ≥ 0.5 (464 transmitted vs. 402 nontransmitted; rate ratio (RR) = 1.15, p-value = 0.038 by binomial test). Over-transmission of rare singleton LGD variants was not observed in genes that are not constrained (RR = 1.03, p-value = 0.31 by binomial test). The excess of transmitted singleton LGD variants in constrained genes increased after removing variants observed in the ExAC database (303 transmitted vs. 242 untransmitted; RR = 1.25, p-value = 0.010 by binomial test). These results provide further evidence that rare, inherited LGD variants in constrained genes confer increased risk for ASD.[14,15] We then searched for known haploinsufficient ASD or neurodevelopmental disorder (NDD) genes (SFARI Gene score 1 or 2 or listed in DDG2P and associated with a neurological phenotype[16]) that are disrupted by the rare singleton LGD variants and are transmitted. We found 13 such variants (2 of them on the X chromosome), as compared with 10 variants that are not transmitted (including one on the X chromosome) (Supplementary Data 3). Manual review of these variants revealed that most of them are not likely pathogenic because they either affect only a subset of transcripts that are not expressed in the majority of tissues,[17] are located close to the 3′ end of the transcript (last 5% of the coding sequence) or are indels that overlap but do not change the sequence of essential splice sites. The results suggest that the rare LGD variants in known ASD/NDD genes have only limited contribution to the overall transmission disequilibrium in this class of variants.

By integrating exome sequence read depth and SNP microarray signal intensity data, we identified 273 rare CNVs (occurring with a carrier frequency of ≤1% of the 1379 individuals in the analysis and also appear <1% in 1000 Genomes population and healthy controls[18]) in 206 affected offspring. Of these, 253 CNVs were inherited (0.544/affected offspring) and were on average 194 kb.

Table 1. Phenotypic description of the 457 families with at least one offspring affected with ASD in the SPARK pilot study

Role	$n = 465$	Average age (years) of ASD diagnosis	Average age (years) at enrollment (SD)	Range of age (years) at enrollment	Intellectual disability (%)	Nonverbal (%)	Epilepsy (%)	ADHD (%)
All offspring with ASD								
Affected male offspring	376	4.8	12.9 (8.4)	1.5–44.6	22% (78/356)	13% (46/356)	7% (25/356)	30% (106/356)
Affected female offspring	89	5.6	12.8 (7.3)	1.9–29.8	33% (28/84)	10% (8/84)	13% (11/84)	23% (19/84)

1379 individuals in 39 multiplex and 418 simplex families were genomically characterized, including 472 individuals (465 offspring and 7 parents) affected with ASD. All phenotypic variables are not available for all participants

These inherited CNVs contained an average of 3.7 genes, which reduces to an average of 0.7 genes that are constrained (pLI ≥ 0.5) (Supplementary Data 4). Similar to the frequency observed in previous studies[8,19] (~5% within a cohort of affected individuals), we identified 20 de novo CNVs (dnCNVs) (0.043/affected offspring) (Supplementary Data 5). On average, dnCNVs were larger (1.6 Mb) and contained more total and constrained genes (19 genes, 5.5 constrained genes with a pLI ≥ 0.5).

Despite the fact we were underpowered to detect statistically significant burden differences between sexes, we observed a trend toward a 1.8-fold higher burden of dnCNVs in ASD females (0.067/female vs 0.037/male, respectively). In contrast, the frequency of rare, inherited CNVs in ASD females and males were similar (0.551/female vs 0.543/male, respectively). Similar to Sanders et al.[8] dnCNVs in female cases also affect more genes than dnCNVs in males (2.3 vs 0.47 genes in dnCNVs per female proband vs per male proband, respectively; p-value = 0.013, Kruskal–Wallis test).

Of the CNVs detected, six mapped to the chromosome 16p11.2 region (three de novo and three inherited in five families). Four of the six 16p11.2 CNVs occurred at the most common breakpoints (BP4-BP5), occurring in 0.9% of affected offspring, consistent with the expected ASD prevalence.[20] Together, the results suggest that the saliva-derived DNA collected in SPARK should provide comparable CNV data to previous studies using DNA derived from whole blood. We also used read-depth and SNP genotypes to identify several chromosomal aneuploidies (Supplementary Fig. 1), including one case of trisomy 21 (47, XY + 21), one case of Klinefelter syndrome (47, XXY), one case of Turner syndrome (45, X), and one case of uniparental iso-disomy of chromosome 6 (UPiD6).

Given their emerging role in genetic risk for ASD and other NDDs, we also assessed postzygotic mosaic mutations[21,22] in the SPARK cohort. In parallel, we utilized a previously established method[23] and a novel approach to identify likely mosaic SNVs (Methods, Supplementary Figs. 3–8). We identified 65 likely mosaic mutations (0.142/offspring) (Supplementary Data 6). The majority of these mutations were unique to the mosaic call set; however, 18 were also identified in the main de novo SNV call set with an average alternative allele fraction of 25.4% (Supplementary Data 6), suggesting that these mutations are likely to have occurred after fertilization. These results indicate that ~10% (65/652) of the total de novo SNVs in the SPARK pilot are of postzygotic origin. Comparing these data to a similar mosaic set from the Simons Simplex Collection (SSC),[23] we found similar mosaic mutation characteristics, despite the fact that different DNA sources, capture reagents, and sequencing instruments were used (Supplementary Fig. 7). Due to the limited number of mosaic calls, we did not attempt to evaluate mosaic mutation burden. However, we observed that a number of potentially mosaic mutations were in known ASD/NDD genes or genes that are constrained (Supplementary Data 6). For example, we identified a potential mosaic LGD variant in *MACF1*, which is highly constrained (pLI = 1), plays essential roles in neurodevelopment, functions through the previously implicated Wnt signaling pathway,[24] and has been recently suggested as a candidate gene based on a dnLGD variant in a Japanese ASD cohort.[25] In *CREBBP*, which reached genome-wide significance in a recent NDD meta-analysis,[16] we identified a potential mosaic missense variant, in addition to two other germline de novo missense variants in SPARK, adding to the evidence that it is an ASD/NDD risk factor. Future work will help determine the contribution of mosaic mutations in such genes to ASD.

Genes with a higher mutational burden

We assessed genes with multiple dnLGD variants in the SPARK cohort and identified four genes with more than one dnLGD variant (*CHD8*, *FOXP1*, *SHANK3*, and *BRSK2*). *BRSK2* is the only gene with multiple dnLGD variants in SPARK that reached genome-wide significance (p-value = 2.3×10^{-6} by one-sided exact Poisson test, <0.05/20,000 genes), although there was one individual in the Autism Sequencing Consortium (ASC) cohort[7] with a dnLGD variant in *BRSK2* (Table 2).

To increase the statistical power to identify new ASD genes, we performed a meta-analysis of de novo variants in 4773 published ASD trios[3,4,7,8] and 465 SPARK trios using TADA[26] (Methods). In this analysis, we included dnLGD variants and de novo D-mis variants, which we defined as those that have a CADD score ≥25.[10] The TADA analysis presumes a model of genetic architecture compatible with the observed burden and recurrence of de novo damaging variants and assigns a false discovery rate (FDR) q-value for each gene based on the number of damaging variants and baseline mutation rates. We identified 67 genes with an FDR threshold of ≤0.1. Of these, there are 26 genes that also harbored a damaging variant in SPARK, most of which are already known ASD/NDD genes. There are six genes (*BRSK2*, *ITSN1*, *PAX5*, *FEZF2*, *DMWD*, and *CPZ*) that reached an FDR threshold of 0.1 only after the inclusion of de novo variants from SPARK (Fig. 1). The association signal for *DMWD* was driven by two LGD variants but the gene is not constrained (pLI = 0), so this gene may be a false positive.

Of the 34 genes listed in Fig. 1, only *BRSK2* meets genome-wide significance as a new ASD risk gene. All four individuals in SPARK, ASC and the SSC with de novo functional variants in *BRSK2* are males with cognitive impairment and severe speech delay (Table 2). *MBD5* and *IRF2BPL* reached an FDR value of ≤0.1 in a previous meta-analysis but not significant in our analysis of published de novo variants,[8] because the previous study also included evidence from de novo CNVs and deleterious variants of unknown inheritance from a case-control sample in that analysis. *MBD5*, as well as *QRICH1*, *SLC6A8*, and *RERE* are known NDD risk genes in the latest DDG2P database.[16]

In our TADA results, we further broadened our focus on genes that harbored damaging variants in the SPARK data and those that had a FDR ≤ 0.2 (Supplementary Data 7). When the TADA analysis is restricted to genes harboring damaging variants in SPARK with an FDR ≤ 0.2, we identified 34 genes (Fig. 1), of which 21 have a known role in ASD or NDDs. We also incorporated inherited variants and CNVs from the SPARK families into the TADA analysis, but did not find additional newly significant genes.

We then searched for additional supporting evidence for a role of these genes in ASD and NDDs, including other deleterious variants in previous studies and case reports not included in the meta-analysis, membership in gene sets previously associated with ASD,[3,4,7,8] and published functional studies (Supplementary Data 8). Recent studies have reported additional individuals with ASD and/or NDD with de novo damaging variants in these genes including *BRSK2*,[27] *PAX5*,[4,28] *NR4A2*,[29,30] *RALGAPB*,[7,31,32] and *DPP6*.[5,33,34]

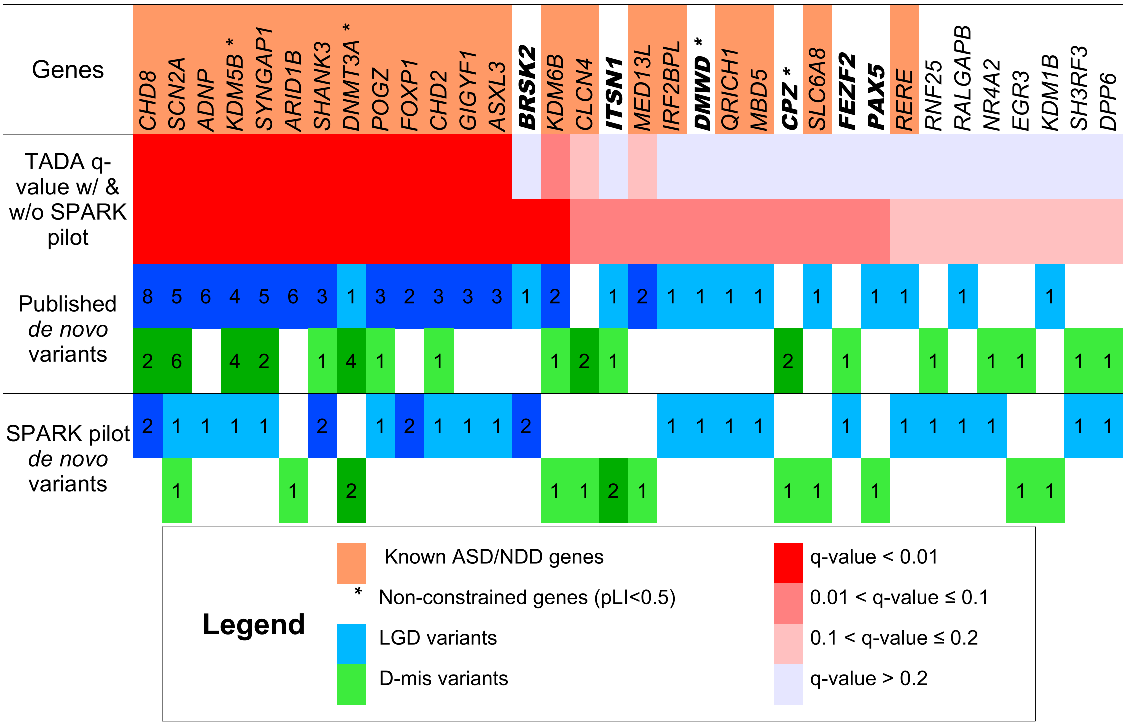
In addition to multiple deleterious variants in these candidate ASD risk genes, we also found evidence that they function in biological pathways previously linked to ASD. For example, mRNA translation of *BRSK2*, *ITSN1*, and *RALGAPB* in neurons is predicted to be regulated by FMR1 protein.[35] In addition, *ITSN1* and *DPP6* are part of the postsynaptic density components in human neocortex.[36] *PAX5* and *FEZF2* are involved in transcription regulation during central nervous system development.[4,24,37] *KDM1B* is a known chromatin modifier, and *EGR3* has been implicated in neurodevelopment.[38,39]

We also searched rare singleton inherited LGD variants of these newly significant genes in SPARK and published SSC data, and

Table 2. Variants in four newly statistically significant and constrained (pLI ≥ 0.5) ASD risk genes (*BRSK2*, *ITSN1*, *FEZF2*, and *PAX5*) in published and SPARK trios and associated phenotypic information

Subject ID	**SP0037695**	**SP0042217**	**08C79336**	**SP0007556**	**SP0025011**	**SP0016887**	**13400.p1**	**13704.p1**	**14637.p1**	**SP0037344**	**11074.p1**	**SP0016232**	**12858.p1**
Cohort	SPARK	SPARK	ASC	SPARK	SPARK	SPARK	SSC	SSC	SSC	SPARK	SSC	SPARK	SSC
Gene	*BRSK2*	*BRSK2*	*BRSK2*	*ITSN1*	*ITSN1*	*ITSN1*	*ITSN1*	*ITSN1*	*ITSN1*	*FEZF2*	*FEZF2*	*PAX5*	*PAX5*
Variant	p.T547fs	c.951-1G > A	c.1365-1G > C	p.P1619L (MPC = 2.03)	p.P156fs	p.Q711X	c.1952 + 1del	p.E576*	p. P164Rfs*22	p.A397fs	p.R344C (MPC = 3.37)	p.E113V (MPC = 2.78)	p.A111fs
Inheritance	de novo	de novo	de novo	de novo	Inherited (paternal)	Inherited (paternal)	Inherited (maternal)	Inherited (paternal)	Inherited (paternal)	de novo	de novo	de novo	de novo
Confirmed by Sanger sequencing?	YES	YES	Not available	YES	YES	YES	Not available	Not available	Not available	YES	YES	YES	YES
Sex	Male	Male	Male	Male	Male	Female	Female	Male	Male	Male	Male	Male	Female
DSM diagnosis	ASD	ASD	ASD	Asperger's disorder	Asperger's disorder	ASD	Autism	PDD-NOS	Autism	ASD	Autism	ASD	Autism
Age at evaluation (years)	8	19	5.3	30	34	2	4.2	4.3	5.4	23	9	14	4
Medical concerns	None	Premature birth (24 weeks), vision/ hearing problems (not specified)	Unknown	Obesity, vision/ hearing problems (not specified)	None	None	Migraines	None	None	None	None	Neurological problems (not specified), sleep disorder	None
Seizures (TRUE/ FALSE)	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Febrile seizures only	FALSE	FALSE
Intellectual Disability (TRUE/ FALSE), IQ (if known)	TRUE	TRUE, ≤25	Unknown	FALSE, ≥129	FALSE	FALSE	FALSE, FSIQ = 116	FALSE FSIQ = 91	TRUE– mild, FSIQ = 63	FALSE	TRUE, FSIQ = 68	TRUE, 55–69	FALSE, FSIQ = 91
Language level (at age of evaluation)	Delayed, single words	Delayed, no words	Delayed	Fluent speech (sentences)	Fluent speech (sentences)	Delayed, single words	Delayed, phrase speech (ADOS mod. 2)	Delayed, phrase speech (ADOS mod. 2)	Delayed, phrase speech (ADOS mod. 2)	Unknown	Fluent speech (sentences; ADOS mod. 3)	Delayed, no words	Delayed, phrase speech (ADOS mod. 2)
Language regression (TRUE/FALSE)	FALSE	FALSE	Unknown	FALSE	Unknown	TRUE	FALSE	TRUE	FALSE	Unknown	TRUE	TRUE	TRUE
Co-morbid psychological diagnoses	Learning disorder, motor skills delay, speech articulation problems, feeding disorder	Motor skills delay, speech articulation problems, feeding disorder, encopresis, enuresis	None	Attention or behavior problems—not specified, mood or anxiety problems—not specified, feeding disorder	Attention or behavior problems—not specified, OCD	Feeding disorder	Anxiety	None	None	None	None	Learning disorder, feeding disorder	None
Early motor delay (TRUE/ FALSE)	FALSE	TRUE	FALSE	FALSE	Unknown	FALSE	FALSE	FALSE	FALSE	Unknown	FALSE	FALSE	FALSE

All damaging variants in SPARK participants within these genes have been confirmed with Sanger sequencing. Damaging variants in *PAX5* and *FEZF2* in the SSC were previously validated.[3] MPC scores are listed for missense mutations. All phenotypic information for SPARK participants was collected online



Genes	🟧 *CHD8*	🟧 *SCN2A*	🟧 *ADNP*	🟧 *KDM5B* *	🟧 *SYNGAP1*	🟧 *ARID1B*	🟧 *SHANK3*	🟧 *DNMT3A* *	🟧 *POGZ*	🟧 *FOXP1*	🟧 *CHD2*	🟧 *GIGYF1*	🟧 *ASXL3*	***BRSK2***	🟧 *KDM6B*	🟧 *CLCN4*	***ITSN1***	🟧 *MED13L*	🟧 *IRF2BPL*	***DMWD*** *	🟧 *QRICH1*	🟧 *MBD5*	***CPZ*** *	🟧 *SLC6A8*	***FEZF2***	***PAX5***	🟧 *RERE*	*RNF25*	*RALGAPB*	*NR4A2*	*EGR3*	*KDM1B*	*SH3RF3*	*DPP6*
TADA q-value w/ & w/o SPARK pilot	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▢	▦	▨	▢	▨	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢	▢
	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▦	▦	▦	▦	▦	▦	▦	▦	▦	▦	▦	▨	▨	▨	▨	▨	▨	▨	▨
Published *de novo* variants	8	5	6	4	5	6	3	1	3	2	3	3	3	1	2		1	2	1	1	1	1		1		1	1		1			1		
	2	6		4	2		1	4	1		1				1	2	1						2		1			1		1	1		1	1
SPARK pilot *de novo* variants	2	1	1	1	1		2		1	2	1	1	1	2					1	1	1	1			1		1	1	1	1			1	1
		1				1		2							1	1	2	1					1	1		1					1	1		

Legend

🟧 Known ASD/NDD genes

* Non-constrained genes (pLI<0.5)

🟦 LGD variants

🟩 D-mis variants

🟥 q-value < 0.01

▦ 0.01 < q-value ≤ 0.1

▨ 0.1 < q-value ≤ 0.2

▢ q-value > 0.2

Fig. 1 Meta-analysis using the TADA framework identifies 34 genes with a false discovery rate (FDR) of ≤0.2. Known ASD genes are defined as those with SFARI Gene[105] score ≤2 or implicated in a previous TADA meta-analysis (FDR ≤ 0.1)[8] and known NDD genes are those listed in the DDG2P database[16] and are colored orange. Deleterious missense (D-mis) variants are defined by CADD score ≥25. A total of 34 genes with at least one de novo damaging variant observed in SPARK pilot trios achieve an FDR ≤ 0.2 after meta-analysis with published trios (total $n = 5238$). Fourteen genes are not classified as known ASD or NDD genes. Six genes (*BRSK2, ITSN1, FEZF2, PAX5, DMWD*, and *CPZ*) that have an FDR ≤ 0.1 only after inclusion of SPARK de novo variants are highlighted. The asterisk symbol indicates genes that are not constrained (pLI < 0.5)

identified five additional cases (three in SSC, two in SPARK) carrying inherited LGD variants of *ITSN1* that likely cause loss of gene function. Interestingly, of the six ASD cases with LGD variants in *ITSN1*, five do not have intellectual disability (Table 2). The less severe phenotype and inheritance from unaffected parents are consistent with the modest effect size, although future studies will help determine if *ITSN1* is a bona fide ASD risk gene. Furthermore, in ASC case-control samples,[7] LGD variants in *ITSN1* were also identified in the controls (three in 5397 ASC controls and comparable with the cumulative AF of 2.5e−4 in gnomAD v2.1), although they were still overrepresented in cases (two in 1601 cases).

Functional network analysis and gene expression patterns in candidate ASD risk genes

To relate the candidate ASD risk genes identified in our TADA analysis to previous knowledge of integrated gene networks in ASD, we scored genes with a TADA FDR ≤ 0.2 and not currently listed in SFARI Gene using forecASD, a new ensemble classifier that integrates spatiotemporal gene expression, heterogeneous network data, and previous gene-level predictors of ASD association.[40] Using forecASD, we derived a single score that ranks the evidence for each gene to be involved in ASD risk. Using this approach, we identified ten genes (*RNF25, DMWD, CLCN4, ITSN1, CPZ, SH3RF3, EGR3, RALGAPB, KDM1B*, and *BRSK2*) that have a TADA FDR ≤ 0.2 and were not listed in the SFARI Gene database. These genes have significantly elevated forecASD scores (p-value = 0.007, Z-test in logistic regression model controlling for contribution of previous TADA scores; Supplementary Fig. 9). Furthermore, two predictive features in forecASD that summarize brain expression support and network support are also found to be significantly elevated over the genome background in the set of these ten genes (p-value = 0.015 and p-value = 0.03, respectively, Wilcoxon test; see Supplementary Fig. 9). Importantly, neither of these metrics uses genetic data directly, so these genes collectively have support across the three independent and distinct domains of genetic, network, and brain expression evidence. These statistical associations are conservative estimates because they compare the distribution of evidence scores among the candidate genes described here to the remainder of the genome, which includes well-established ASD genes. Eight of these genes, *BRSK2, KDM1B, RALGAPB, EGR3, SH3RF3, CPZ, ITSN1*, and *CLCN4* fall in the top decile of forecASD scores (the top decile being a recommended cutoff used to define probable ASD risk genes), supporting these genes as having similar properties overall compared with known ASD risk genes.

To illustrate the network context of these eight candidate ASD risk genes, we clustered them along with genes scoring within the top decile of forecASD (Fig. 2a). Network analysis yielded ten tightly connected clusters with distinct biological functions (Supplementary Data 9). Several genes were assigned to clusters that showed enrichment for gene sets consistent with their known functions, including *KDM1B*,[41] *BRSK2*,[42] and *ITSN1*[43] consistent with published functional evidence (Supplementary Data 8). In a subsequent analysis, the interactions between known and novel ASD candidate risk genes were visualized (Fig. 2b). This subnetwork was significantly interconnected (p-value = 5.0×10^{-179} by hypergeometric test), with novel genes showing significantly more functional associations with known ASD candidate risk genes than expected by chance (p-value = 0.005 by hypergeometric test).

Using coexpression networks seeded by high-confidence ASD risk genes, a previous study found that cortical projection neurons

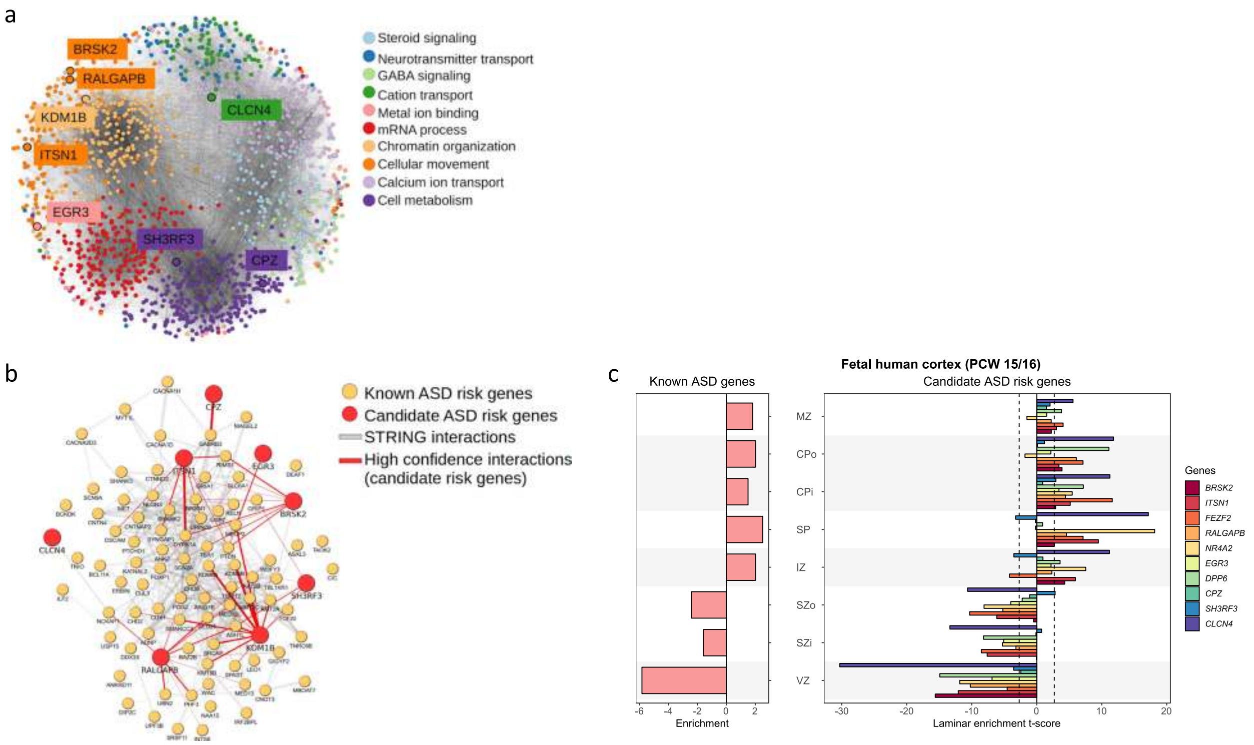


Fig. 2 Network analysis and gene expression of candidate ASD risk genes. **a** STRING networks of forecASD genes, **b** STRING networks of known ASD genes, and **c** gene expression of human fetal cortex at postconceptual weeks (PCW) 15–16. Known ASD genes are defined as those with a SFARI Gene scores[104] ≤2 (84 genes, indicated as SFARI) or implicated in a previous TADA meta-analysis[8] at an FDR ≤ 0.1 (65 genes, indicated as TADA). The enrichment for each gene was measured by the *t*-statistics comparing the expression level in each layer against all other layers. The enrichment of a gene set is the mean of *t*-statistics of its genes. Two candidate ASD risk genes (*PAX5* and *KDM1B*) are not shown due to the low expression levels in human developing cortex (RPKM <1 for at least 20% available neocortical samples in BrainSpan[48]). Data were extracted from Supplementary Tables of Parikshak et al.[45] Laminae abbreviations: marginal zone (MZ), outer/inner cortical plate (CPo/CPi), subplate (SP), intermediate zone (IZ), outer/inner subventricular zone (SZo/SZi), ventricular zone (VZ)

in layers V and VI of the human midfetal prefrontal and primary motor-somatosensory cortex are a key point of convergence for ASD risk genes.[44] Another study also showed that unbiased gene co-expression networks overrepresented with candidate ASD risk genes are more highly expressed in the cortical plate and subplate laminae of the developing human cortex, which will go on to form mature layers II–VI of the cerebral cortex.[45] One of the newly statistically significant genes we identified, *FEZF2*, is a powerful master regulator gene critical for establishing corticospinal neurons,[46] which connect layer Vb of the cortex to the spinal cord, and is known to be expressed in the putative layer V in the late mid-fetal human cortex.[47]

We evaluated gene expression of the candidate ASD risk genes identified by either the TADA meta-analysis and forecASD with regard to cortical layer specificity in the human developing brain.[48] Ten of these genes (*BRSK2, ITSN1, FEZF2, RALGAPB, NR4A2, EGR3*, *DPP6, CPZ, SH3RF3*, and *CLCN4*) have expression data in developing fetal human cortex, and similar to Parikshak et al.[45] they show a trend of increased expression at postconceptual week (PCW) 15–16 (Fig. 2c) and PCW 21 (Supplementary Fig. 10) in the cortical plate and subplate laminae, which will form layers II–VI of the mature cerebral cortex. The mean of *t*-statistics of these ten genes in the inner cortical plate (CPi) and subplate (SP) are greater than two standard deviations (SD) from the mean of randomly selected genes matched for gene length and GC content ($P < 0.01$ by simulation).

We further evaluated cell-type specificity using recently published single-cell RNA-seq data from fetal and adult mouse and human brains[49] (Supplementary Figs. 11 and 12), and found the expression specificity of these candidate ASD risk genes is highest in pyramidal neurons in the mouse hippocampus CA1 region with an enrichment of 3.4 SD from the bootstrapped mean (p-value = 9.6e−3 by simulations controlling gene length and GC content, Supplementary Fig. 11). The enrichment in pyramidal neurons is also observed in the hippocampus CA1 region in human (2.4 SD above the bootstrapped mean, p-value = 0.02 by simulation) using recently published human single nucleus RNA-seq data.[50] These results are consistent with a previous study showing that ASD protein–protein interaction networks related to the 16p11.2 CNV display significantly enriched expression during mid-fetal development as well as early childhood in cerebral cortex.[51] Taken together, we find that the candidate ASD risk genes identified in this study demonstrate differential expression patterns similar to that of known ASD risk genes, providing further support that these genes function in similar biological pathways and mechanisms as known ASD risk genes.

Diagnostic yield in SPARK

Families in the pilot study were selected without regard to genetic diagnosis. Thirteen of the 457 families self-reported a genetic diagnosis, and all were confirmed by our analyses and serve as positive controls to validate our genomic analyses (Supplementary Data 10). For the remaining 444 families, we identified 50 (10.4%) deleterious genetic variants (8 dnCNVs, 14 inherited CNVs, 23 de novo SNVs or indels, 3 inherited LGD variants and 2 chromosomal aneuploidies) in known ASD risk genes or loci in 49 affected individuals (Supplementary Data 10). We also identified an additional 19 likely deleterious genetic variants (1 dnCNV, 1 inherited CNV, 14 de novo SNVs and 3 inherited SNVs) in possible

ASD risk genes or loci in an additional 14 individuals (3.4%). For all cases, we defined deleterious ASD-associated variants as those meeting likely pathogenic or pathogenic criteria according to ACMG standards.[52] We defined possible ASD-associated variants as either SNVs that are de novo missense variants that affect known NDD or ASD genes and have an MPC score[11] ≥2, loss-of-function variants that disrupt possible NDD or ASD genes, or CNVs that delete one or more possible NDD or ASD genes or duplicate known ASD or NDD loci. Possible ASD genes include those that are newly significant in this study (FDR ≤ 0.2) with independent evidence from literature, or genes implicated by other studies with multiple LGD variants found in affected ASD and NDD cases (summarized Supplementary Data 8). We did not search for or discover any incidental findings unrelated to ASD in these families. When DNA was available, ASD-associated genetic findings were confirmed by Sanger sequencing or chromosome microarray, and genetic results were returned to the families ($n = 28$).

DISCUSSION

Overall, the genomic characterization of 457 ASD families (418 simplex and 39 multiplex) in SPARK implicates a number of candidate risk genes in ASD that converge on similar biological networks as known ASD risk genes. We identified a returnable genetic result related to ASD in 10.4% of affected offspring and have begun returning individual genetic results to the families after confirming results in a clinical laboratory. Not surprisingly, our diagnostic yield was highest in affected individuals who also report presence of seizures (27%). The yield in individuals who also report intellectual disability was also higher (20%) than the overall cohort.

In our analysis, our diagnostic yield in affected offspring in multiplex families (15.2%) was slightly higher than affected offspring in simplex families (10.1%). Interestingly, the genetic findings in multiplex families rarely explained ASD in all affected family members (Supplementary Fig. 13), similar to previous studies that have also found affected siblings with discordant mutations.[53,54] For example, in a family with an affected father and three affected children, the most severely affected child harbored a dnLGD in *ADNP*. No other family member carried this variant or any other identifiable contributing variant. In another pedigree, an affected male child with an affected father inherited a 15q11.2 BP1-BP2 deletion from a mother who does not report an ASD diagnosis, but we found no contributing variant in the affected father. We also identified eight families in which there was greater than one contributing variant, even in families in which we were unable to identify contributing variants in all affected offspring. In one family with two affected children, the female child inherited a 1q21.1 CNV from an unaffected mother and also harbored a dnLGD in *RALGAPB*. However, the affected male child did not harbor either of these variants and the CNV identified, like many potentially pathogenic variants, is known to be variably expressive. Future studies with larger sample sizes will allow for a more robust comparison of the genetic architecture of ASD in simplex vs. multiplex families.

Over time, we expect the diagnostic yield in SPARK to increase as more individuals with ASD are studied and as additional genetic risk factors are identified. For example, we identified LGD variants in *MEIS2* and *AKAP10* and deletions of the *NFIB, DLL1*, and *HNRNPD* genes. Although these genes did not reach statistical significance in our TADA meta-analysis, their role in ASD is supported by multiple mutations in the literature, and they likely represent other candidate ASD risk genes (Supplementary Data 8). We interpreted those variants as possible contributors to ASD in those individuals. The genetic findings in those cases will be confirmed and returned in the future if and when these genes are established as ASD risk genes.

Using a systems biology approach, we demonstrated that the newly statistically significant and candidate ASD risk genes identified in this analysis are well-supported beyond genetic association and are predicted to be ASD risk genes based on a variety of functional properties, including patterns of spatiotemporal gene expression in the brain and protein network connectivity. *BRSK2* and seven of the candidate ASD risk genes scored in the top decile of forecASD, an integrator of published functional evidence for ASD risk genes (Supplementary Fig. 9). The genes localized to network clusters representing processes critical for neurodevelopment (Fig. 2), including chromatin modification (*KDM1B*), neuronal polarity[55] (*BRSK2*), and neuronal migration of pyramidal neurons[56] (*ITSN1*). The candidate ASD risk genes also showed significant over-connectivity to known ASD risk genes (*p*-value = 0.005 by hypergeometric test, Fig. 2b). Together, the TADA genetic association analysis coupled with the supporting functional and network-level data triangulate these genes as being robust and biologically plausible contributors to ASD risk.

Despite the limited sample size in this pilot study, we were able to identify four newly statistically significant ASD genes. Power analysis using a simulation-based approach confirmed that the observed yield is expected given the presumed genetic architecture in the TADA analysis (Supplementary Table 1). We expect to identify ~70–75% of all ASD risk genes in the future that meet a similar FDR threshold (0.1–0.2) when we reach SPARK's goal of sequencing 50,000 complete trios (Supplementary Table 1). Other analyses of large cohorts in ASD are underway, including a recent analysis of ~12,000 individuals with ASD.[57] This study, which used a mixture of family-based and case-control data, found statistical support for 99 ASD risk genes, increasing the number of ASD risk genes from 65.[8] Future meta-analyses of both SPARK data and other ASD cohort data are planned to maximize ASD risk gene discovery.

For many genes identified with de novo damaging variants, inherited loss-of-function variants in affected individuals were not found (Kosmicki et al. [15] and this study), suggesting our current knowledge about ASD risk genes is biased toward those with high penetrance. Future studies with larger sample sizes will be needed to identify and validate additional risk genes of lower penetrance that confer inherited ASD risk.

Altogether, these data suggest that the methods used to ascertain individuals with ASD, saliva collection, and genomic data are of high quality, and future analysis of the tens of thousands of families enrolling in SPARK will significantly contribute to our understanding of the genetic basis of ASD. By returning genetic results to participants, we expect to increase engagement and increase the number of recontactable participants for genetically targeted clinical research and trials.

METHODS

Participant recruitment, phenotyping, and DNA sequencing

All participants were recruited to SPARK under a centralized IRB protocol (Western IRB Protocol #20151664). All participants provided written informed consent to take part in the study. Written informed consent was obtained from all legal guardians or parents for all participants age 18 and younger and all participants age 18 and older who have a legal guardian. Assent was also obtained from dependent participants age 10 and older. Participants are asked to fill out questionnaires online as described here: https://www.sfari.org/spark-phenotypic-measures/. Families are classified as multiplex if the initial individual with ASD registered in the study has a first-degree family member with ASD, as indicated either by enrollment or survey report.

Essential phenotypic information was curated across language and motor development, co-morbidities, and Repetitive Behavior Scale-Revised,[58] Social Communication Questionnaire-Lifetime[59] and Developmental Coordination Disorder Questionnaire score[60] (Table 2). In SSC, all phenotype details were determined through clinic evaluation and interview; specifically, language delay was defined by Autism Diagnostic

Observation Schedule module (1–4) per age,[61] and regression was determined from the Autism Diagnostic Interview-Revised.[62] For SPARK, all variables were taken from parent report. It was noted that rates of language disorder and psychiatric co-morbidities are lower in SSC likely due to DSM-IV diagnostic practice at the time.

Saliva was collected using the OGD-500 kit (DNA Genotek) and DNA was extracted in a CLIA-certified laboratory at the Baylor Miraca Genetics Laboratories (Houston, TX) or PreventionGenetics (Marshfield, WI). Exome capture was performed using VCRome and the spike-in probe set PKv2 at the Baylor College of Medicine Human Genome Sequencing Center (Houston, TX). Captured exome libraries were sequenced using the Illumina HiSeq platform in 100 bp paired end reads. Samples were sequenced to a minimum standard of >85% of target covered at 20×, and on average, 96% of the target was sequenced to 20×. The Illumina HumanCoreExome (550K SNP sites) array was used for genotyping.

Read alignment and QC

Postsequencing reads were aligned to build 37 of the human genome using bwa version 0.6.2-r126,[63] duplicates were marked using Picard version 1.93 MarkDuplicates, and indels were realigned using GATK[64] version 2.5-2-gf57256b IndelRealigner. Quality checks were performed on the BAM files using SAMTools[65] version 1.3.1 flagstat and Picard version 2.5.0 CalculateHsMetrics. Overall, 98 ± 1.8% of the reads mapped to the genome, 96 ± 2.3% of the reads were properly paired reads, and 87 ± 15% of targeted regions had ≥10× coverage.

KING[66] was used for relatedness inference based on the genotype of exome SNPs (MAF >0.01). Estimated kinship coefficient and number of SNPs with zero shared alleles (IBS0) between a pair of individuals were plotted. Parent–offspring, sibling pairs, and unrelated pairs can be distinguished as separate clusters on the scatterplot (Supplementary Fig 1). One outlier parent–offspring pair (SP0002452 and mother) showed higher than expected IBS0 and was caused by parental chr6 iso-UPD. Pairwise scatterplots of heterozygotes to homozygotes (het/hom) ratio of chromosome X, sequencing depth of chromosome X and Y normalized by the mean depth of autosomes were used for sex check. Two samples with sex chromosome aneuploidy were identified as outliers in the scatterplot (Supplementary Fig. 2).

Variant calling

De novo SNV/indel detection. De novo sequence variants were called by three groups—University of Washington (UW), Simons Foundation (SF), Columbia University Medical Center (CUMC)—according to the methods below.

UW. Variants were called from whole exome sequence (WES) using FreeBayes[67] and GATK.[64] FreeBayes version v1.1.0-3-g961e5f3 was used with the following parameters: –use-best-n-alleles 4 -C 2 -m 20 -q 20; and GATK version 3.7 HaplotypeCaller was used with the following parameters: -A AlleleBalanceBySample -A DepthPerAlleleBySample -A MappingQualityZeroBySample -A StrandBiasBySample -A Coverage -A FisherStrand -A HaplotypeScore -A MappingQualityRankSumTest -A MappingQualityZero -A QualByDepth -A RMSMappingQuality -A ReadPosRankSumTest -A VariantType. Postcalling bcftools[68] version 1.3.1 norm was used with the following parameters -c e -O z -s -m –both. We identified candidate de novo calls based on the intersection of FreeBayes and GATK VCF files and identifying variants present in offspring but not in parents. We required a minimum of ten sequence reads in all members of the parent–offspring trio; an allele balance >0.25 and a PHRED quality >20 for both FreeBayes and GATK variants.

SF. Sequence data were preprocessed using GATK best practices and variant calls were predicted using three variant callers: GATK v3.6,[69] FreeBayes v1.1.0-441, and Platypus v0.8.1-0.[70] GATK: gVCF files were generated for each sample with GATK HaplotypeCaller (minimum confidence thresholds for calling and emitting was set to 30 and 10, respectively); joint variant calls were performed using GATK GenotypeGVCFs with the recommended default hard filters. For SNPs, we filtered out: QD <2.0 || FS >60.0 || MQ <40.0 || MQRankSum <−12.5 || ReadPosRankSum <−8.0. For indels, we filtered out: QD <2.0 || FS >200.0 || ReadPosRankSum <−20.0. FreeBayes: variants were called with default settings for optimal genotyping of indels in lower-complexity sequence. The final data set included candidate calls with a quality of 5 or greater. Platypus: variant calling was performed with local assembly analysis when at most ten haplotypes were allowed. Variants were filtered out for allele bias (p-value < 0.0001), bad reads (>0.9), sequence complexity (>0.99) and RMSMQ (<20); other filters were applied on estimated haplotype population frequency (FR), total coverage at the locus (TC) and phred-scaled quality of reference allele (QUAL): (FR[0] < = 0.5 and TC < 4 and QUAL < 20),or (TC < 13 and QUAL < 10),or (FR[0] > 0.5 and TC < 4 and QUAL < 50). For each variant caller, a variant was identified as a candidate de novo variant if the variant was called in the proband and it occurred only once in the cohort, with an alternative allele fraction between 0.2 and 0.8. Both parents were required to have the homozygous reference genotype at the de novo locus. Read coverage of the variant locus had to be at least ten reads in each sample in the trio. De novo candidate variants were classified by DNMFilter algorithm[71] that was retrained with the SSC data set[3,14]: 1800 de novo mutations identified by both Iossifov et al.[3] and Krumm et al.,[14] 1104 validated SNVs and indels from both studies and 400 variants that failed validation. We also randomly selected ~3000 negative examples from the pool of all SSC variants that were not confirmed to be de novo. After merging de novo candidate variants from three variant callers, candidate de novos were considered if they occurred only once in the cohort, passed hard filters, and had assigned de novo probability greater than 0.88 for SNVs and greater than 0.0045 for small indels. In the latter case, the total parental alternative allele count <3 reads.

CUMC. Variants were called from aligned sequence data using GATK HaplotyperCaller to generate individual level gVCF files. All samples in the cohort were then jointly genotyped and have variant quality recalibrated by GATK v3.8.[64] A variant present in the offspring with homozygous reference genotypes in both parents was considered to be a potential de novo variant. We used a series of filters to identify de novo variants. Briefly, we included variants that passed VQSR filter (tranche ≤ 99.7 for SNVs and ≤99.0 for indels) and had GATK's Fisher Strand ≤ 25, quality by depth ≥2. We required the candidate de novo variants in probands to have ≥5 reads supporting the alternative allele, ≥20% alternative allele fraction, Phred-scaled genotype likelihood ≥60 (GQ), and population AF ≤0.1% in ExAC; and required both parents to have ≥10 reference reads, <5% alternative allele fraction, and GQ ≥ 30.

De novo SNV/indel consensus call set and annotation. De novo variants were independently called by three centers—UW, SF, CUMC. De novo variants called by all three groups were included in the final list by default. Those called by one or two groups were manually evaluated and included in the final list if consensus was reached among all groups after discussion and manual inspection with IGV plots. Variants were annotated by ANNOVAR[72] based on GENCODE Basic v19.[73] Candidate variants in the ACMG secondary findings v2 59 gene list[74] (except *PTEN*, *TSC1*, and *TSC2*) were excluded. Coding de novo variants—nonsense, missense, or synonymous SNVs, frameshift or nonframeshift indels, and splicing site variants—were annotated. De novo variants were also annotated with snpEff version 4.1g[75] (reference GRCh37.75), SFARI Gene scores (version q1, 2018, https://gene.sfari.org/database/gene-scoring/), CADD,[10] MPC[11] and findings from Deciphering Developmental Disorders project (gene2phenotype).[16]

Inherited singleton variants. We first performed following filtering on individual genotypes. We required minimal read-depth ≥10 and GQ ≥30, required allelic balance <0.1 for homozygotes reference, >0.9 for homozygotes alternative, and 0.3–0.7 for heterozygotes SNVs (0.25–0.75 for heterozygotes indels). Genotype calls not passing those criteria were set to missing. Then we removed variants having missing genotypes in >25% of founders. We focused analysis on singleton variants in which the alternative allele was only seen in one parent in the data. We calibrated GATK's VQS LOD score for SNV and indels separately such that synonymous singleton SNVs and nonframeshift singleton indels were transmitted 50% of the time (Supplementary Fig. 14) The resulting VQS LOD score cutoffs are −1.85 for SNVs and −1.51 for indels. As mentioned in the Results section, inherited LGD variants are less likely to cause a complete loss of function to the gene. To prioritize inherited LGD variants, we require the variant to be annotated as HC (high-confidence) by LOFTEE v0.3[12] using default parameters in >60% of the GENCODE transcripts.

Identification of mosaic mutations. Mosaic SNVs were independently called by two centers—Oregon Health & Science University (OHSU) and CUMC. The OHSU approach was previously published[23] and utilized a binomial deviation and logistic regression model to score candidate mosaic variants. The CUMC approach used a novel approach that was based on a beta-binomial deviation and an FDR based approach to determine per site thresholds.

OHSU. SNVs were called as previously described.[23] In brief, pileups were generated using SAMtools (v 1.1) with BAQ disabled and mapQ 29 (*samtools mpileup –B –q 29 –d 1500*) on processed BAMs. Variants were called on individual samples using VarScan 2.3.2, LoFreq 2.1.1 and an in-house mpileup parsing script (mPUP). Additional parameters for Varscan included: *–min-var-freq* 1×10^{-15} *–p-value 0.1*. Per sample caller outputs were combined and annotated using ANNOVAR (03/22/15 release) with databases: Refseq genes (obtained 03/2017), segmental duplications (UCSC track genomicSuperDups, obtained 03/25/2015), repetitive regions (UCSC track simpleRepeat and hg19_rmsk, obtained 03/25/2015), Exome Aggregation Consortium (ExAC) release 0.3 (obtained 11/29/2015), Exome Sequencing Project (ESP) 6500 (obtained 12/22/2014), and 1000 Genomes Phase 3 version 5 (obtained 12-16-2014).

Variants were filtered based on the best practices established in Krupp et al.:[23] (1) variant must be exonic or disrupt a canonical splice site, (2) have a population frequency of ≤0.5%, (3) have at least five alternative reads, (4) not be in a known segmental duplication or repetitive regions (SDTRF), (5) called by at least two variant callers, (6) SPARK cohort count ≤1 and SSC cohort count ≤2, (7) variant read mismatch ≤3, and (8) allele fraction upper 90% confidence interval ≤0.05. For a variant to be considered de novo, parental alternative allele count must be ≤4 reads. De novo variants were considered to be candidate mosaic variants if: (1) the probability the allele fraction significantly deviated from heterozygous (PHET) was ≤0.001, (2) the allele fraction upper 90% confidence interval was <0.4, and (3) a logistic regression model score was ≥0.518.

CUMC. SNVs were called on a per-trio basis using SAMtools (v1.3.1-42) and BCFtools (v1.3.1–174). We generated trio VCF files using samtools '*mpileup*' command with options '*–q 20 –Q 13*' corresponding to mapQ and baseQ thresholds of 20 and 13 respectively, followed by bcftools '*call*' with option '*–p 1.1*' to expand the set of variant positions to be evaluated for mosaicism. In contrast to the OHSU pipeline, BAQ was used to potentially reduce false positive SNV calls caused by misalignments.[76] To identify de novo variants from trio VCF files, we selected for sites with (i) a minimum of six reads supporting the alternate allele in the proband and (ii) for parents, a minimum depth of ten reads and 0 alternate allele read support. Variants were then annotated using ANNOVAR (v2017-07-17) to include information from refGene, gnomAD (March 2017), 1000 Genomes (August 2015), ExAC, genomicSuperDups, COSMIC (v70), and dbSNP (v147) databases. CADD,[10] MPC[11] were used to annotate variant functional consequence.

Preprocessing and QC. To reduce the noise introduced by our variant calling approach, we preprocessed our variants using a set of filters. Since our method is allelic depth-dependent, we took a conservative filtering approach to reduce the impact of false positives on model parameter estimation. We first filtered our variant call set for rare heterozygous coding variants (MAF $\leq 1\times10^{-4}$ across all populations represented in gnomAD and ExAC databases). To account for regions in the reference genome that are more challenging to resolve, we removed variant sites found in regions of nonunique mappability (score <1; 300 bp), likely segmental duplication (score >0.95), and known low complexity.[77] We then excluded sites located in *MUC* and *HLA* genes and imposed a maximum variant read depth threshold of 500. To account for common technical artifacts, we used SAMtools PV4 *p*-values with a threshold of 1×10^{-3} to exclude sites with evidence of baseQ bias, mapQ bias, and tail distance bias. To account for potential strand bias, we used an in-house script to flag sites that have either (1) 0 alternate allele read support on either the forward or reverse strand or (2) $p < 1\times10^{-3}$ and OR < 0.33 or OR > 3 when applying Fisher's method to compare strand based reference or alternative allele counts. Finally, we excluded sites with frequency >1% in the SPARK pilot, as well as sites belonging to outlier samples (with abnormally high de novo SNV counts, cutoff = 7) and complex variants (defined as sites with neighboring de novo SNVs within 10 bp).

IGV visualization of low allele fraction de novo SNVs. To identify likely false positives among our low allele fraction (VAF <0.3) de novo SNVs, we used Integrative Genomics Viewer (IGV v2.3.97) to visualize the local read pileup at each variant across all members of a given trio. We focused on the allele fraction range 0.0–0.3 since this range captures the majority of the technical artifacts that will negatively impact downstream parameter estimation. Sites were filtered out if (1) there were inconsistent mismatches in the reads supporting the mosaic allele, (2) the site overlapped or was adjacent to an indel, (3) the site had low MAPQ or was not primary alignment, (4) there was evidence of technical bias (strand, read position, tail distance), or (5) the site was mainly supported by soft-clipped reads.

Empirical bayes postzygotic mutation detection model. To distinguish variant sites that show evidence of mosaicism from germline heterozygous sites, we modeled the number of reads supporting the variant allele (N_{alt}) as a function of the total site depth (N). In the typical case, N_{alt} follows a binomial model with parameters N = site depth and *p-value* = mean VAF. However, we observed notable overdispersion[78,79] in the distribution of variant allele fraction compared with the expectations under this binomial model. To account for this overdispersion, we instead modeled N_{alt} using a beta-binomial distribution. We estimated an overdispersion parameter θ for our model whereby for site depth values N in the range 1–500, we (1) bin variants by identifying all sites with depth N, (2) calculate a maximum-likelihood estimate θ value using N and all N_{alt} values for variants in a given bin, and (3) estimate a global θ value by taking the average of θ values across all bins, weighted by the number of variants in each bin.

We used an expectation-maximization (EM) algorithm to jointly estimate the fraction of mosaics among apparent de novo mutations and the FDR of candidate mosaics. This initial mosaic fraction estimate gives a prior probability of mosaicism independent of sequencing depth or variant caller and allows us to calculate, for each variant in our input set, the posterior odds that a given site is mosaic rather than germline.

Finalized union mosaic call set and validation selection: The high confidence call sets from the two parallel mosaic determination approaches were combined, and all candidate mosaic variants were then inspected manually in IGV. Variants in regions with multiple mismatches or poor mapping quality were removed, and the remaining mosaics comprised the high confidence mosaic call set. For calls that were unique to one approach, the variant was annotated with which quality filter it initially failed. Variants that were flagged as low confidence germline by CUMC approach but mosaic by OHSU approach had posterior odds >1 and were thus retained in the union call set.

CNV detection: De novo and rare inherited CNVs were independently called by two centers—UW and SF. The final CNV list included all autosomal CNVs that were called by both SF and UW pipelines either with reciprocal overlap of at least 50% or when the CNV from one pipeline was completely within the CNV from the other pipeline. In both cases, the overlapping region was reported as the final region and annotated as described below. CNVs called only by one pipeline were considered as high confidence CNVs if they were called by at least two tools or if they were de novo CNVs confirmed by manual inspection of plots on exome data. High confidence CNVs were also included in the final list after discussion and manual inspection of plots on exome data. De novo CNVs were additionally inspected on BAF and LRR plots on genotyping data. CNVs that had at least 75% overlap with known segmental duplications (segDups track for hg19 from UCSC browser) were excluded. All CNVs were annotated with the list of RefSeq HG19 genes, OMIM genes, brain embryonically expressed genes,[3] brain critical genes,[19] ASD significant,[80] and ASD related genes[8,14] that have their coding regions overlapping with the CNV. CNVs greater than or equal to 50 kbp in size were annotated with morbidity map[81] case and control frequencies using a 50% reciprocal overlap while CNVs < 50 kbp were annotated with their frequency in the 1000 genomes project[82] using a 50% reciprocal overlap. We do note that it is possible some events may be missed with this annotation because of different platforms (e.g. exome, array, and genome), but the two analyses provide reasonable insight into the population prevalence of large and smaller CNVs in the general population. In addition, each found gene was annotated with pLI (ExAC release 0.3, http://exac.broadinstitute.org/downloads), ASD,[83] RVIS,[84] LGD,[85] and SFARI Gene scores (version q1, 2018, https://gene.sfari.org/database/gene-scoring/). dnCNVs that affect DUSP22 and olfactory genes were excluded due to high variability in copy number of those regions among individuals.[86]

UW, detection using XHMM and CoNIFER. CNVs from WES were called using CoNIFER and[87] XHMM.[88] CoNIFER version v0.2.2 was used with the *S* value, –svd 7, set as a threshold as suggested by the scree plot. XHMM version statgen-xhmm-3c57d886bc96 was used with the following parameters –minTargetSize 10 –maxTargetSize 10000 –minMeanTargetRD 10 –maxMeanTargetRD 500 –minMeanSampleRD 25 –maxMeanSampleRD 200 –maxSdSampleRD 150 to filter samples and targets, and then to mean-center the targets; PVE_mean –PVE_mean_factor 0.7 was used to normalize mean-centered data using PCA information; –maxSdTargetRD 30 was used to filter and z-score centers (by sample) the PCA normalized data; and then to discover CNVs in all samples. Calls from CoNIFER and XHMM were merged in a VCF file using https://github.com/zeeev/mergeSVcallers with the following parameters -t xhmm,conifer -r 0.5 -s 50000, then merged VCF was sorted by Picard version v2.5.0, and zipped and indexed with Tabix version v0.2.6. We re-genotyped each XHMM and CoNIFER CNV event by assessing the RPKM values from the CoNIFER workflow on an individual. Probands were considered to have a deletion if their average RPKM value was less than −1.5 s.d and have a duplication if their average RPKM value was greater than 1.5 s.d. For an event to be considered as variant in a parent, we required an average ZRPKM less than −1.3 or greater than 1.3 for deletions and duplications, respectively.

UW, CNV validation using SNP microarray. We generated an independent CNV callset for validation purpose using SNP microarray genotyping data generated from Illumina InfiniumCoreExome-24_v1.1, where IDATs ($n =$ 1,421) were processed using Illumina Genome Studio Software. CNV analysis was performed using the Illumina CNVpartition algorithm version v3.2.0. Log R Ratio data for all samples and probes was exported. PennCNV[89] version v1.0.4 was used to detect CNVs with the following parameters -test –hmm -pfb all.pfb –gcmodelfile –confidence. We determined the maximum and minimum overlap of SNP microarray CNVs based on the presence of WES probes to make the array calls more similar to the exome calls and considered an event to have support by PennCNV or CNVpartition if there was at least 50% reciprocal overlap. We also generated per probe copy number estimates using CRLMM[90,91] version 1.38.0 as previously described[14] and genotyped each candidate WES CNV. Deletions were considered variant if they had a p-value less than 0.05 and a mean percentile rank less than 30. Duplications were considered variant if they had a p-value less than 0.05 and a percentile rank of mean greater than 70. CNVs passing the RPKM genotyping were combined with the CNV data from CRLMM, PennCNV, and CNVPartition. We considered WES CNVs as valid if there was support for gain or loss from the PennCNV, CNVpartition, or CRLMM approaches described above. We assessed inheritance using both SNP and WES data and preferentially scored inherited events over de novo CNVs.

SF. CNVs were called with two tools - xHMM v 1.0[92] and CLAMMS v 1.1.[93] xHMM**:** CNVs were called with default settings (except not filtering on the maximum target size), including filtering low complexity and GC extreme targets. CLAMMS: CNVs were called with INSERT_SIZE = 390 bp and training per-sample-models on sample specific reference panels due to the observed batch effect in the data; CLAMMS calls were filtered for all CNVs with Q_EXACT less than 0, or Q_SOME less than 100, or CNVs that were in samples with more than 70 predicted CNVs of the size at least 10 Kb and of quality score Q_SCORE at least 300. The inheritance status of the autosomal CNVs was determined by default xHMM protocol for de novo CNVs identification with plink 1.07[94] and Plink/Seq 0.10 [https://atgu.mgh.harvard.edu/plinkseq/]. Similar protocol was implemented in java for CLAMMS analysis. For each tool, two tiers of CNV calls—the most confident calls (tier 1) and less confident calls (tier 2)—were defined, based on de novo and transmission rates for different cuts on quality scores: SQ (phred-scaled quality of some CNV event in the interval) and NQ (phred-scaled quality of not being diploid, i.e., DEL or DUP event in the interval) in xHMM and Q_SOME (phred-scaled quality of any CNV being in this interval) in CLAMMS. xHMM tier1 included all autosomal CNVs with both SQ and NQ quality scores of at least 60, and tier2—all autosomal CNV calls with quality scores between 30 and 60. Samples with more than 10 de novo CNVs in xHMM tier1 of size at least 10 kb were excluded. CLAMMS tier1 included all predictions with quality score 999, except predictions for 25 probands that have CNVs of size greater than 500 kb with quality score 999 or predictions, which region was partially inherited and partially de novo; tier 2 included those excluded from tier 1 predictions as well as all CNVs with quality score Q_SOME at least 400 and less than 999. Predictions by both methods that had less than 3 exons or at least 75% overlap with known segmental duplications (segDups track for hg19 from UCSC browser) were removed from the list. The final list of CNV predictions included all CNVs from tier 1 predicted by either xHMM or CLAMMS and "intersection" of tier 2 sets from both tools, that is, CNVs that were confirmed by two tools with reciprocal or cumulative reciprocal overlap of at least 50%. In the latter case, CNV predicted by one tool is covered by a set of CNVs predicted by the other tool. If a CNV from xHMM or CLAMMS was confirmed by the other tool, the overlapping region was reported as the final region. CNVs were removed from the analysis if it had more than half of its length overlapping with the ACMG secondary findings v2 gene[74] (except *PTEN, TSC1*, and *TSC2*). If such gene covers less than 50% of CNV, the part of CNV without the gene was kept if it has at least 25% of its length not covered by segmental duplications. To identify higher confidence CNV predictions, xHMM and CLAMMS plots were manually investigated for each CNV in the final SF list. In addition, SF predictions were compared with PennCNV[89,95,96] calls from array data, which have confidence score of at least 100. All reciprocal overlaps of at least 50% were treated as additional evidence for CNV support.

UW, chromosome aneuploidy assessment. We also assessed evidence of chromosomal aneuploidy by calculating sequence read depth using SAMTools[10] version 1.4 on a per chromosome basis normalizing by the relative density of WES probes and comparing the normalized value for each chromosome to the normalized value on chromosome 1 (assumed to be diploid). For autosomes, we multiplied this number by two to get the estimate of chromosomal copy number. We did not multiply by two for the X or Y chromosomes. To further assess the chromosomal copy number, the heterozygosity was calculated for all SNPs and indels. For heterozygous sites, the absolute mean deviation from 0.5 was also calculated. We assessed both metrics to identify outliers. Aneuploidies were required to have support from both the read depth and SNP/indel metrics.

Burden of de novo variants

Baseline mutation rates for different classes of de novo variants in each GENCODE coding gene were calculated using a previously described mutation model.[9] Briefly, the trinucleotide sequencec context was used to determine the probability of each base mutating to each other possible base. Then the mutation rate of each functional class of point mutations in a gene was calculated by adding up the mutation rate of each nucleotide in the longest transcript. The rate of frameshift indels was presumed to be 1.1 times the rate of nonsense point mutations. The expected number of variants in different gene sets were calculated by summing up the class-specific variant rate in each gene in the gene set multiplied by twice the number of patients (and if on chromosome X, further adjusted for female-to-male ratio[97]).

The observed number of variants in each gene set and case group was then compared with the baseline expectation using a Poisson test. In all analyses, constrained genes were defined by a pLI score of ≥0.5. To compare with previously published ASD studies, we collected published de novo variants identified in 4773 simplex trios from three largest ASD studies to date.[3,4,7] To account for platform differences, the baseline mutation rate of each gene was scaled so that the exome-wide expected number of silent variants matches the observed count.

TADA analysis

To perform TADA analysis of de novo variants, we assumed the fraction of disease genes is 5% as estimated by previous studies.[26,98] The prior relative risk for LGD variants and D-mis (defined by CADD > = 25) were specified as Gamma (18,1) and Gamma (6,1). The prior mean relative risks were determined using the relationship between burden and relative risk as described previously.[26] The baseline mutation rate of each gene was the same as used in burden analysis. The analysis was performed on de novo variants of 4773 published trios and after combing de novo variants identified from SPARK pilot trios.

Laminal layer and cell type enrichment

To evaluate the expression specificity of laminal layer of human developing cortex, we analyzed RNA-seq data of neocortical samples of BrainSpan[48] following the method of Parikshak et al.[45] The expression specificity was measured by a t-statistic comparing the expression level in each layer against all other layers. Two candidate ASD risk genes (*PAX5, KDM1B*) were not included in the analysis due to the low expression levels (RPKM <1 for at least 20% available neocortical samples). To evaluate cell-type specificity, we used published data of mouse neuronal cell types inferred from analyzing single cell RNA-seq data of fetal and adult mouse brains generated by the Karolinska Institutet (KI),[99] and human CNS cell types inferred from a single nucleus RNA-seq data.[50] The mouse orthologs of human genes were retrieved from MGI database.[100] The cell-type specificity was measured by a specificity index which is the mean expression level in one cell type over the summation of mean expression level across all cell types.[101] To analyze the overall trend of specificity of a gene set, the mean specificity measure of its genes was compared with

10,000 sets of randomly drawn genes matched for the transcript length and GC content and the enrichment is measured by the standard deviation from the mean specificity of random gene sets.[101]

Network and functional analysis

The network depicted in Fig. 2a was constructed using the top decile of forecASD genes, SFARI Genes scoring 1 or 2, and SPARK newly implicated genes (6 in total). These genes were projected onto the STRING network[102] (v10) using the igraph R package (1708 genes). Edges within the STRING network were thresholded at 0.4, according to the authors' recommendation. The largest connected subcomponent (1664 genes) was then extracted as the basis for further network analysis. Clustering was performed on the fully connected network using the fastgreedy community function available within the igraph package. Clusters with fewer than 30 genes were not considered for further analysis (none of these clusters contained the six genes highlighted here). Following the first round of clustering, clusters with >150 genes were subject to an additional round of clustering, with the goal of separating broad functions of genes into more specific subcomponents. This process resulted in ten clusters. Each cluster was assessed for functional enrichment using the Gene Ontology[103] as accessed through the clusterProfiler package within R. During the functional analysis the background gene universe was always set to the full set of genes represented among the ten clusters. Visualization of this network analysis was performed in Cytoscape.[104] The top five most significant GO terms associated with each cluster are available in the Supplementary Data 9. Cluster labels in Fig. 2 were chosen as the most representative among the top terms for each cluster. Figure 2b was constructed using the subset of the larger network (Fig. 2a), corresponding to SPARK newly implicated genes and SFARI Gene genes scoring 1 or 2 (88 genes). These genes were projected onto the STRING network within Cytoscape using the STRINGapp. All nonzero-weighted edges were considered. The fully connected component was visualized, which resulted in two genes being dropped (*DEAF1* and *RANBP17*). Edges adjacent to newly implicated genes with a STRING interaction score of ≥0.4 are highlighted.

ForecASD analysis. We used a recently developed method, forecASD[40] that indexes support for a gene being related to ASD by integrating genetic, expression, and network evidence through machine learning. We examined the forecASD scores of candidate ASD risk genes from the TADA analysis and compared them to the remainder of the genome using a Wilcoxon rank-sum test. We similarly used the Wilcoxon test and employed two predictive features used by forecASD (BrainSpan_score and STRING_score) to assess whether the new genes showed similarity to known ASD risk genes in terms of brain expression patterns and network connectivity. Importantly, because forecASD uses previously published TADA scores among its predictive features, which are strongly correlated with updated TADA scores, we investigated whether the elevated forecASD scores in our candidate genes could be explained solely by the previous TADA scores. Specifically, we fit a logistic regression model with the candidate ASD risk genes labeled as '1' and 500 size-matched background genes (not listed in the SFARI gene database) labeled as '0' in the dependent variable (Y). Separate models were fit using either forecASD or TADA[8] scores as predictors, or both together in a full model. Both TADA and forecASD were significantly associated with the "new gene" indicator when considered in isolation ($P < 0.001$ for both, Z-test on logistic regression coefficients). However, when included together in a model of Y, forecASD remained significantly associated (p-value = 0.00012, Z-test on logistic regression coefficients) while TADA lost significance (p-value = 0.41, Z-test on logistic regression coefficients). The Akaike information criterion (AIC) indicated that the forecASD-only model was a more optimal fit compared with either the TADA-only or TADA + forecASD fit. This analysis suggests that the elevated forecASD scores observed in the ten new genes cannot be fully explained by the use of TADA as a predictor in forecASD.

Reporting summary

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

CODE AVAILABILITY

Methods for SNV, Indels, CNV analysis are available at https://genomicpipelines.sparkforautism.org/.

DATA AVAILABILITY

The genomic and phenotypic data for the 1379 individuals analyzed in this study is available by request from SFARIBase (https://www.sfari.org/resource/sfari-base/) with accession ID: SFARI_SPARK_WES_p.

ACKNOWLEDGEMENTS

We are extremely grateful to the thousands of individuals and families who are participating in this study. We are grateful to the many ASD advocacy and service organizations that have helped us inform the community about SPARK, including the Autism Society of America and its affiliates, Autism Speaks, Autism Science Foundation, Easter Seals, Arkansas Autism Resource and Outreach Center, Global and Regional Asperger's Syndrome Partnership, Kentucky Autism Training Center, and Autistic Self Advocacy Network. We thank the members of SPARK's Community Advisory Council for providing feedback and advice. We thank members of our Scientific and Community Advisory Board and SFARI scientists for advice on our protocol and participant outreach and retention strategies. We thank PreventionGenetics for managing and processing biospecimens, DNA Genotek for handling saliva kit logistics, and Baylor College of Medicine Human Genome Sequencing Center for exome sequencing. The SPARK initiative is funded by the Simons Foundation as part of SFARI. This research was supported, in part, by a grant from the National Institute of Mental Health (NIMH R01MH101221) and a grant from the Simons Foundation (SFARI #608045) to E.E.E., a grant from the National Institute of Mental Health to T.N.T. (1K99MH117165) and grants MH105527 and DC014489 from the National Institute of Health to L.B. and J.J.M. E.E.E. is an investigator of the Howard Hughes Medical Institute. B.J.O. is a Klingenstein-Simons Fellow (Esther A. & Joseph Klingenstein Fund, Simons Foundation).

AUTHOR CONTRIBUTIONS

P.F., X.Z., I. Astrovskaya, T.N.T., J.J.M., B.J.O., N.V., Y.S. and W.K.C. designed and conceived the study. A.C., A.C.G., A.D.S., A.E., A.G., A.J., A.J.A., A.L.R., A.M., A.M.D., A.N., A.N.S., A.P., A.P. M., A.R.S., A. Swanson, B.A.H., B.E.R., B. Hauf, B.J.O., B.M.V., B.V., C.A., C.A.E., C.C., C.E.R., C. Harkins, C. Hayes, C.H.W., C.J.S., C. Lord, C.O., C.R.R., C.T., D.E.S., D.G.A., D.I., D. Lee Coury, D. Li, E.A.F., E. Berry-Kravis, E.C., E.J.F., E.L., E.L.W., E.M.B., E.O., E.T.M., G.M., G.S.D., H.E.K., H.H., H. Lam Schneider, H. Lechniak, H. Li, H. Zaydens, I. Arriaga, J.A., J.A.G., J.F.C., J.G., J.L., J.M., J.N., J.O., J.P., J.P., J.S., J.S.S., J.T., J.T.M., J. Wallace, K.A.S., K.C., K.E.H., K.G.P., K.L.P., K.O., K. Roeder, L.A., L.A.C., L. Beeson, L.D., L.D.P., L.G.S., L.M.H., L.M.P., L.P.G., L.S., L.V.S., L. Wasserburg, L. Casey White, L.Y.H., M.A., M.C., M. Heyman, M. Jones, M. Jordy, M.J.M., M.N.H., M.S., M.T., M.Y., N.B., N. Hanna, N. Harris, N. Lillie, N.R., N.T., O.Y.O., P.F., P.M., R.A. Bernier, R.D.A., R.D.C., R.J.L., R.P.G., R. Remington, R.T.S., S.B., S.C., S.E., S.F., S.G., S.H., S.J., S.J.B., S.J.L., S.L.F., S.L.H., S. M.K., S.P., S.Q., S. Sandhu, S.T., S.W., V.J.M., V.S., W.K.C., W.S.Y., and Z.W. recruited participants and collected clinical data and biospecimens. A. Balasubramanian, A. Beaudet, A.F., A.H., A.J. Griswold, A.K., A. Soucy, B.J.O., C.L Martin, C.N., D.H.G., E. Berry-Kravis, E. Bahl, E.E.E., E.R., H. Doddapaneni, H.Q., H.Zhang, H.Zhao, I. Astrovskaya, J.B.H., J. H., J.J.M., J.S.S., J.V., J. Wang, K. Rajbhandari, L. Brueggeman, L.G.S., L.P.G., L.Z., M.G., M.R.G., M.Y.D., N.V., P.F., R.A. Barnard, R.D., R.D.A., S.L., S.M., S. Santangelo, S. Skinner, S.X., T.K., T.N. T., T.P., T.S., T.T., T.W., T.Y., V.G., V.K., W.K.C., X.L., X.Z., Y.D., Y.S., and Z.M. helped with data interpretation. A.A., A. Bashar, A.E.L., A. Salomatov, A.S.C., B. Han, C.M.S., C.R., I. Astrovskaya, I.F., J.A., J.B.H., K.L., M.D.M., M.E.B., M.K., N.C., N. Lawson, N.V., R.M., R. Rana, S. G., S. Shah, S.X., W.C., and W.J. built and supported the SPARKforAutism.org website, software, databases and systems, and managed SPARK data. A.D.K., A.H., A.J. Gruber, A. Sabo, B.J.O., C.E., D.M., E. Brooks, G.J.F., I. Astrovskaya, J.J.M., L. Brueggeman, L.G.S. M.P., M. R.G., N.V., P.F., R.A. Barnard, R.A.G., T.N.T., T.W., W.K.C., X.Z., and Y.S. performed analyses, processed biospecimens and sequenced DNA samples. P.F. and W.K.C. supervised the work. P.F., X.Z., I. Astrovskaya, T.N.T., J.J.M., B.J.O., N.V., W.K.C., and Y.S. wrote this paper.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Genomic Medicine* website (https://doi.org/10.1038/s41525-019-0093-8).

Competing interests: M.S. has received research funding from Roche, Novartis, Pfizer, Aucta, Navitor, Rugen, Ibsen, Neuren, LAM Therapeutics, Quadrant Biosciences and has served on the Scientific Advisory Board of Sage Therapeutics, Roche and Takeda. D.H.G. receives research funding from Takeda Pharmaceuticals, and consulting fees or equity participation for scientific advisory board work from Ovid Therapeutics, Axial Bio-therapeutics, Acurastem, and Falcon Computing. E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. All other authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Schendel, D. E. et al. Association of psychiatric and neurologic comorbidity with mortality among persons with autism spectrum disorder in a danish population. *JAMA Pedia*. **170**, 243–250 (2016).
2. Sandin, S. et al. The heritability of autism spectrum disorder. *JAMA* **318**, 1182–1184 (2017).
3. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
4. RK, C. Y. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
5. Tammimies, K. et al. Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. *JAMA* **314**, 895–903 (2015).
6. SPARK Consortium SPARK: A US Cohort of 50,000 families to accelerate autism research. *Neuron* **97**, 488–493 (2018).
7. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209–215 (2014).
8. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
9. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
10. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet*. **46**, 310–315 (2014).
11. Samocha, K. E. et al. Regional missense constraint improves variant deleteriousness prediction. https://doi.org/10.1101/148353 (2017).
12. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
13. Jacquemont, S. et al. A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet*. **94**, 415–425 (2014).
14. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
15. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet*. **49**, 504–510 (2017).
16. Wright, C. F. et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet* **385**, 1305–1314 (2015).
17. Consortium, G. T., et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
18. Cooper, G. M. et al. A copy number variation morbidity map of developmental delay. *Nat. Genet*. **43**, 838–846 (2011).
19. Uddin, M. et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nat. Genet*. **46**, 742–747 (2014).
20. Weiss, L. A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
21. Freed, D. & Pevsner, J. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet*. **12**, e1006245 (2016).
22. Lim, E. T. et al. Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.* **20**, 1217–1224 (2017).
23. Krupp, D. R. et al. Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am. J. Hum. Genet*. **101**, 369–390 (2017).
24. Chen, H. J. et al. The role of microtubule actin cross-linking factor 1 (MACF1) in the Wnt signaling pathway. *Genes Dev.* **20**, 1933–1945 (2006).
25. Takata, A. et al. Integrative analyses of de novo mutations provide deeper biological insights into autism spectrum disorder. *Cell Rep.* **22**, 734–747 (2018).
26. 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).
27. Hiatt, S. M. et al. Deleterious Variation in BRSK2 Associates with a Neurodevelopmental Disorder. *Am. J. Hum. Genet*. **104**, 701–708 (2019).
28. O'Roak, B. J. et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. *Nat. Commun.* **5**, 5595 (2014).
29. Reuter, M. S. et al. Haploinsufficiency of NR4A2 is associated with a neurodevelopmental phenotype with prominent language impairment. *Am. J. Med Genet. A* **173**, 2231–2234 (2017).
30. Levy, J. et al. NR4A2 haploinsufficiency is associated with intellectual disability and autism spectrum disorder. *Clin. Genet*. **94**, 264–268 (2018).
31. Epi, K. C. et al. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
32. Guo, H. et al. Inherited and multiple de novo mutations in autism risk genes suggests a multifactorial model. *Mol. Autism* **9**, 64 (2018).
33. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
34. Maussion, G. et al. Implication of LRRC4C and DPP6 in neurodevelopmental disorders. *Am. J. Med Genet. A* **173**, 395–406 (2017).
35. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
36. Bayes, A. et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* **14**, 19–21 (2011).
37. Han, W. et al. TBR1 directly represses Fezf2 to control the laminar origin and development of the corticospinal tract. *Proc. Natl Acad. Sci. USA* **108**, 3041–3046 (2011).
38. Tourtellotte, W. G. & Milbrandt, J. Sensory ataxia and muscle spindle agenesis in mice lacking the transcription factor Egr3. *Nat. Genet*. **20**, 87–91 (1998).
39. Li, L. et al. Egr3, a synaptic activity regulated transcription factor that is essential for learning and memory. *Mol. Cell Neurosci.* **35**, 76–88 (2007).
40. Brueggeman, L., Koomar, T. & Michaelson, J. Forecasting autism gene discovery with machine learning and genome-scale data. https://doi.org/10.1101/370601 (2018).
41. Stukenbrock, E. H. et al. KdmB, a jumonji histone H3 demethylase, regulates genome-wide H3K4 trimethylation and is required for normal induction of secondary metabolism in aspergillus nidulans. *PLOS Genet*. **12**, https://doi.org/10.1371/journal.pgen.1006222 (2016).
42. Lilley, B. N., Pan, Y. A. & Sanes, J. R. SAD kinases sculpt axonal arbors of sensory neurons through long- and short-term responses to neurotrophin signals. *Neuron* **79**, 39–53 (2013).
43. Gryaznova, T., Gubar, O., Burdyniuk, M., Kropyvko, S. & Rynditch, A. WIP/ITSN1 complex is involved in cellular vesicle trafficking and formation of filopodia-like protrusions. *Gene* **674**, 49–56 (2018).
44. Willsey, A. J. et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
45. Parikshak, N. N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
46. Lodato, S. et al. Gene co-regulation by Fezf2 selects neurotransmitter identity and connectivity of corticospinal neurons. *Nat. Neurosci.* **17**, 1046–1054 (2014).
47. Gu, Z. et al. Control of species-dependent cortico-motoneuronal connections underlying manual dexterity. *Science* **357**, 400–404 (2017).
48. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
49. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet*. **50**, 825–833 (2018).
50. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
51. Lin, G. N. et al. Spatiotemporal 16p11.2 protein network implicates cortical late mid-fetal brain development and KCTD13-Cul3-RhoA pathway in psychiatric diseases. *Neuron* **85**, 742–754 (2015).
52. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
53. Yuen, R. K. et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–191 (2015).
54. Leppa, V. M. et al. Rare Inherited and De Novo CNVs Reveal Complex Contributions to ASD Risk in Multiplex Families. *Am. J. Hum. Genet*. **99**, 540–554 (2016).
55. Kishi, M., Pan, Y. A., Crump, J. G. & Sanes, J. R. Mammalian SAD kinases are required for neuronal polarization. *Science* **307**, 929–932 (2005).
56. Jakob, B. et al. Intersectin 1 is a component of the Reelin pathway to regulate neuronal migration and synaptic plasticity in the hippocampus. *Proc. Natl Acad. Sci. USA* **114**, 5533–5538 (2017).
57. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. (2019). https://www.biorxiv.org/content/10.1101/484113v3.
58. Bodfish, J. W., Symons, F. J., Parker, D. E. & Lewis, M. H. Varieties of repetitive behavior in autism: comparisons to mental retardation. *J. Autism Dev. Disord.* **30**, 237–243 (2000).
59. Rutter, M., Bailey, A. J. & Lord, C. Social communication questionnaire (SCQ). (2003).
60. Wilson, B. N. et al. Psychometric properties of the revised Developmental Coordination Disorder Questionnaire. *Phys. Occup. Ther. Pedia.* **29**, 182–202 (2009).
61. Lord, C. et al. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).
62. Lord, C., Rutter, M. & Le Couteur, A. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Dev. Disord.* **24**, 659–685 (1994).

63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
64. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
65. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
67. Garrison, E., Gabor, M. Haplotype based variant detection from short-read sequencing. *arXiv* https://arxiv.org/abs/1207.3907 (2012).
68. Danecek, P. & McCarthy, S. A. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics* **33**, 2037–2039 (2017).
69. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
70. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
71. Liu, Y., Li, B., Tan, R., Zhu, X. & Wang, Y. A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. *Bioinformatics* **30**, 1830–1836 (2014).
72. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
73. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
74. Kalia, S. S. et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SFv2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* **19**, 249–255 (2017).
75. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
76. Li, H. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157–1158 (2011).
77. Li, W. & Freudenberg, J. Mappability and read length. *Front. Genet.* **5**, 381 (2014).
78. Heinrich, V. et al. The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.* **40**, 2426–2431 (2012).
79. Ramu, P. et al. Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. *Theor. Appl. Genet.* **126**, 2051–2064 (2013).
80. Stessman, H. A. et al. Targeted sequencing identifies 91 neurodevelopmental-disorder risk genes with autism and developmental-disability biases. *Nat. Genet.* **49**, 515–526 (2017).
81. Coe, B. P. et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).
82. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
83. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
84. Petrovski, S. et al. GenicIntolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics* **9** (2013).
85. Iossifov, I. et al. Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl Acad. Sci. USA* **112**, E5600–E5607 (2015).
86. Young, J. M. et al. Extensive copy-number variation of the human olfactory receptor gene family. *Am. J. Hum. Genet.* **83**, 228–242 (2008).
87. Krumm, N. et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res* **22**, 1525–1532 (2012).
88. Fromer, M. et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
89. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665–1674 (2007).
90. Ritchie, M. E., Carvalho, B. S., Hetrick, K. N., Tavare, S. & Irizarry, R. A. R/Bioconductor software for Illumina's Infinium whole-genome genotyping BeadChips. *Bioinformatics* **25**, 2621–2623 (2009).
91. Scharpf, R. B., Irizarry, R. A., Ritchie, M. E., Carvalho, B. & Ruczinski, I. Using the R Package crlmm for genotyping and copy number estimation. *J. Stat. Softw.* **40**, 1–32 (2011).
92. Fromer, M. & Purcell, S. M. Using XHMM software to detect copy number variation in whole-exome sequencing data. *Curr. Protoc. Hum. Genet* **81**, 7 23 21–21 (2014).
93. Packer, J. S. et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics* **32**, 133–135 (2016).
94. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
95. Diskin, S. J. et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
96. Wang, K. et al. Modeling genetic inheritance of copy number variations. *Nucleic Acids Res.* **36**, e138 (2008).
97. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
98. Nguyen, H. T. et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* **9**, 114 (2017).
99. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 e1022 (2018).
100. Eppig, J. T. et al. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res.* **43**, D726–D736 (2015).
101. Skene, N. G. & Grant, S. G. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* **10**, 16 (2016).
102. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
103. Ashburner, M. et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.* **25**, 25–29 (2000).
104. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
105. Larsen, E. et al. A systematic variant annotation approach for ranking genes associated with autism spectrum disorders. *Mol. Autism* **7**, 44 (2016).

THE SPARK CONSORTIUM

Leonard Abbeduto[9], John Acampado[1], Andrea J. Ace[1], Charles Albright[10], Michael Alessandri[11], David G. Amaral[9], Alpha Amatya[1], Robert D. Annett[12], Ivette Arriaga[13], Ethan Bahl[4], Adithya Balasubramanian[14], Nicole Bardett[15], Asif Bashar[1], Arthur Beaudet[14], Landon Beeson[16], Raphael A. Bernier[17], Elizabeth Berry-Kravis[18,19,20], Stephanie Booker[21], Stephanie J. Brewster[22], Elizabeth Brooks[1], Martin E. Butler[1], Eric M. Butter[10], Kristen Callahan[12], Alexies Camba[13], Sarah Carpenter[15], Nicholas Carriero[1], Lindsey A. Cartner[1], Ahmad S. Chatha[1], Wubin Chin[1], Renee D. Clark[23], Cheryl Cohen[24], Eric Courchesne[25], Joseph F. Cubells[26], Mary Hannah Currin[23], Amy M. Daniels[1], Lindsey DeMarco[27], Megan Y. Dennis[28], Gabriel S. Dichter[23], Yan Ding[14], Huyen Dinh[14], Ryan Doan[29], HarshaVardhan Doddapaneni[14], Sara Eldred[10], Christine Eng[14], Craig A. Erickson[21], Amy Esler[30], Ali Fatemi[31], Gregory J. Fischer[32], Ian Fisk[1], Eric J. Fombonne[33], Emily A. Fox[17], Sunday Francis[34], Sandra L. Friedman[35], Swami Ganesan[1], Michael Garrett[36], Vahid Gazestani[37], Madeleine R. Geisheker[3], Jennifer A. Gerdts[17], Daniel H. Geschwind[38], Robin P. Goin-Kochel[39], Anthony J. Griswold[40], Luke P. Grosvenor[1], Angela J. Gruber[32], Amanda C. Gulsrud[13], Jaclyn Gunderson[34], Anibal Gutierrez[11], Melissa N. Hale[11], Monica Haley[13], Jacob B. Hall[1], Kira E. Hamer[17], Bing Han[1], Nathan Hanna[10], Christina Harkins[21], Nina Harris[15], Brenda Hauf[23], Caitlin Hayes[10], Susan L. Hepburn[41], Lynette M. Herbert[11], Michelle Heyman[42], Brittani A. Phillips[9], Susannah Horner[27], Jianhong Hu[14], Lark Y. Huang-Storms[33], Hanna Hutter[31], Dalia Istephanous[34], Suma Jacob[34], William Jensen[1], Mark Jones[10], Michelle Jordy[23], A. Pablo Juarez[43], Stephen Kanne[44,45], Hannah E. Kaplan[25], Matt Kent[1], Alex Kitaygorodsky[2], Tanner Koomar[4], Viktoriya Korchina[14], Anthony D. Krentz[32], Hoa Lam Schneider[11], Elena Lamarche[23], Rebecca J. Landa[31], Alex E. Lash[1], J. Kiely Law[24], Noah Lawson[1], Kevin Layman[1], Holly Lechniak[16], Sandra Lee[14], Soo J. Lee[16], Daniel Lee Coury[10], Christa Lese Martin[46], Deana Li[9], Hai Li[19], Natasha Lillie[34], Xiuping Liu[14], Catherine Lord[47], Malcolm D. Mallardi[1], Patricia Manning[21], Julie Manoharan[1], Richard Marini[1], Gabriela Marzano[39], Andrew Mason[48], Emily T. Matthews[49], James T. McCracken[13], Alexander P. McKenzie[31], Zeineen Momin[14], Michael J. Morrier[26], Shwetha Murali[3], Vincent J. Myers[1], Jason Neely[31], Caitlin Nessner[14], Amy Nicholson[43], Kaela O'Brien[21], Eirene O'Connor[1], Cesar Ochoa-Lubinoff[19], Jessica Orobio[39], Opal Y. Ousley[26], Lillian D. Pacheco[33], Juhi Pandey[27], Anna Marie Paolicelli[42], Katherine G. Pawlowski[49], Karen L. Pierce[25], Joseph Piven[22], Samantha Plate[27], Marc Popp[32], Tiziano Pramparo[25], Lisa M. Prock[49], Hongjian Qi[2], Shanping Qiu[42], Angela L. Rachubinski[35], Kshitij Rajbhandari[14], Rishiraj Rana[1], Rick Remington[1], Catherine E. Rice[26], Chris Rigby[1], Beverly E. Robertson[1], Katherine Roeder[26], Cordelia R. Rosenberg[35], Nicole Russo-Ponsaran[50], Elizabeth Ruzzo[38], Mustafa Sahin[51], Andrei Salomatov[1], Sophia Sandhu[13], Susan Santangelo[52], Dustin E. Sarver[12], Jessica Scherr[10], Robert T. Schultz[27], Kathryn A. Schweers[16], Swapnil Shah[1], Tamim Shaikh[35], Amanda D. Shocklee[45], Laura Simon[34], Andrea R. Simon[39], Vini Singh[31], Steve Skinner[53], Kaitlin Smith[9], Christopher J. Smith[48], Latha V. Soorya[16], Aubrie Soucy[29], Alexandra N. Stephens[1], Colleen M. Stock[1], James S. Sutcliffe[54], Amy Swanson[15], Maira Tafolla[13], Nicole Takahashi[45], Taylor Thomas[4], Carrie Thomas[21], Samantha Thompson[9], Jennifer Tjernagel[1], Bonnie Van Metre[31], Jeremy Veenstra-Vanderweele[55], Brianna M. Vernoia[1], Jermel Wallace[26], Corrie H. Walston[23], Jiayao Wang[2], Zachary Warren[43], Lucy Wasserburg[34], Loran Casey White[1], Sabrina White[12], Ericka L. Wodka[31], Simon Xu[1], Wha S. Yang[13], Meredith Yinger[26], Timothy Yu[29], Lan Zang[14], Hana Zaydens[1], Haicang Zhang[2] and Haoquan Zhao[2]

[9]MIND Institute and Department of Psychiatry and Behavioral Sciences, University of California, Davis, Sacramento, CA 95817, USA; [10]Division of Pediatric Psychology and Neuropsychology, Nationwide Children's Hospital, Columbus, Ohio 43205, USA; [11]Department of Psychology, University of Miami, Coral Gables, FL 33146, USA; [12]Department of Pediatrics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [13]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA 90095, USA; [14]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [15]Vanderbilt Kennedy Center, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [16]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL 60612, USA; [17]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA 98195, USA; [18]Department of Neurological Sciences, Rush University Medical Center, Chicago, IL 60612, USA; [19]Department of Pediatrics, Rush University Medical Center, Chicago, IL 60612, USA; [20]Department of Biochemistry, Rush University Medical Center, Chicago, IL 60612, USA; [21]Department of Psychiatry and Behavioral Neuroscience, Cincinnati Children's Hospital Medical Center - Research Foundation, Cincinnati, OH 45229, USA; [22]Translational Neuroscience Center, Boston Children's Hospital, Boston, MA 02115, USA; [23]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; [24]Hugo W. Moser Research Institute at Kennedy Krieger, Baltimore, MD 21211, USA; [25]Department of Neurosciences, University of California, San Diego, School of Medicine, La Jolla, CA 92037, USA; [26]Department of Psychiatry and Behavioral Sciences, Emory Autism Center, Emory University, Atlanta, GA 30033, USA; [27]Center for Autism Research, Children's Hospital of Philadelphia, Philadelphia, PA 19146, USA; [28]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, Sacramento, CA 95616, USA; [29]Department of Medicine, Boston Childrens Hospital, Boston, MA 02115, USA; [30]Department of Pediatrics, University of Minnesota, Minneapolis, MN 55414, USA; [31]Center for Autism and Related Disorders, Kennedy Krieger Institute, Baltimore, MD 21211, USA; [32]PreventionGenetics, Marshfield, WI 54449, USA; [33]Departments of Pediatrics & Psychiatry, Oregon Health & Science University, Portland, OR 97239, USA; [34]Department of Psychiatry, University of Minnesota, Minneapolis, MN 55455, USA; [35]Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO 80045, USA; [36]Department of Pharmacology and Toxicology, University of Mississippi Medical Center, Jackson, MS 39216, USA; [37]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92161, USA; [38]Department of Neurology, University of California, Los Angeles, CA 90095, USA; [39]Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA; [40]John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL 33136, USA; [41]Human Development & Family Studies, Colorado State University, Fort Collins, CO 80524, USA; [42]Department of Psychiatry, Weill Cornell Medicine, White Plains, New York 10605, USA; [43]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [44]Department of Health Psychology, University of Missouri, Columbia, MO 65211, USA; [45]Thompson Center for Autism and Neurodevelopmental Disorders, University of Missouri, Columbia, MO 65211, USA; [46]Geisinger Autism & Developmental Medicine Institute, Lewisburg, PA 17837, USA; [47]David Geffen School of Medicine, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA 90024, USA; [48]Research, Southwest Autism Research and Resource Center, Phoenix, AZ 85006, USA; [49]Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA; [50]Department of Behavioral Sciences, Rush University Medical Center, Chicago, IL 60612, USA; [51]Department of Neurology, Boston Children's Hospital, Boston, MA 02115, USA; [52]Maine Medical Center Research Institute, Portland, ME 04101, USA; [53]Greenwood Genetic Center, Greenwood, SC 29646, USA; [54]Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN 37232, USA and [55]Department of Psychiatry, Columbia University Medical Center, New York, NY 10032, USA