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

Pamela Feliciano1, Xueya Zhou2, Irina Astrosvkaya3, Tychele N. Turner4, Tianyun Wang3, Leo Brueggeman4, Rebecca Barnard5, Alexander Hsieh6, LeeAnne Green Snyder1, Donna M. Muzny6, Aniko Sabo6, The SPARK Consortium, Richard A. Gibbs6, Evan E. Eichler7,77, Jacob J. Michaelson2, Irina Astrovskaya4, Natalia Volfovsky1, Yufeng Shen2 and Wendy K. Chung1,8

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.

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.

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 lifelong 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.1,4 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
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 model9 (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 (Supplementary Data 2) and show a 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 (singleton) 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 controls18) in 206 affected offspring. Of these, 253 CNVs were inherited (0.544/offspring) and were on average 194 kb.
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, 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. 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, XXX), one case of Turner syndrome (45, X), and one case of uniparental iso-dismy of chromosome 6 (UPiD6).

Given their emerging role in genetic risk for ASD and other NDDs, we also assessed postzygotic mosaic mutations in the SPARK cohort. In parallel, we utilized a previously established method 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), 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, and has been recently suggested as a candidate gene based on a dnLGD variant in a Japanese ASD cohort. In CREBBP, which reached genome-wide significance in a recent NDD meta-analysis, 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 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 and 465 SPARK trios using TADA (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. The TADA analysis provides 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 CP2) that reached an FDR threshold of 0.1 only after the inclusion of de novo variants from SPARK (Fig. 1). The association between CP2 and 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). MBDS and IRF2BP1 reached an FDR value of ≤0.1 in a previous meta-analysis but not significant in our analysis of published de novo variants, 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. MBDS, as well as QRICH1, SLC6A8, and REE are known NDD risk genes in the latest DDG2P database.

In our TADA results, we further broadened our focus on genes that harbored damaging variants in the SPARK data and those that had an 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, 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, PAX5, NAX4A2, RALGAPB, RERE, and DPP6.

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. In addition, ITSN1 and DPP6 are part of the postsynaptic density components in human neocortex, and PAX5 and FEZF2 are involved in transcription regulation during central nervous system development. A known chromatin modifier, and EGR3 has been implicated in neurodevelopment.

We also searched rare singleton inherited LGD variants of these newly significant genes in SPARK and published SSC data, and
| Subject ID     | SP0037695 | SP0042217 | 08C79336 | SP0007556 | SP0025011 | SP0016887 | 13400.p1 | 13704.p1 | 14637.p1 | SP0037344 | 11074.p1 | SP0016232 | 12858.p1 |
|---------------|------------|------------|----------|------------|------------|------------|----------|----------|----------|------------|----------|------------|----------|
| Gene          | BRSK2      | BRSK2      | AS2C     | BRSK2      | BRSK2      | ITSN1      | ITSN1    | ITSN1    | ITSN1    | PAX5       | ITSN1    | PAX5       | PAX5     |
| Variant       | p.T547fs   | c.951-1G > A | c.1365-1G > C | p.P1619L (MPC = 2.03) | p.P156fs   | p.Q711X    | p.E576*  | p.E113V  | p.E113V  | p.A397fs   | p.E576*  | p.A111fs   | p.A111fs |
| Inheritance   | de novo    | de novo    | de novo  | de novo    | de novo    | Inherited  | Inherited | Inherited | Inherited | Inherited  | Inherited | Inherited  | Inherited |
| Confirmed by Sanger sequencing? | YES       | YES       | Not available | YES       | YES       | Not available | NOT available | Not available | Not available | YES | YES       | YES |
| Sex           | Male       | Male       | Male     | Male       | Male       | Male       | Female   | Male     | Female   | Male       | Male     | Male       | Male     |
| DSM diagnosis | ASD        | ASD        | ASD      | Male       | Male       | ASD        | Autism   | Autism   | ASD      | Autism     | ASD      | Autism     | Autism   |
| Age at evaluation (years) | 8         | 19         | 5.3      | 30         | 34         | 2          | 42       | 43       | 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 | None | Migraines | None | None | None | None | Neurological problems (not specified), sleep disorder |
| Seizures (TRUE/FALSE) | FALSE     | TRUE       | FALSE   | FALSE      | FALSE      | FALSE      | FALSE    | FALSE    | FALSE    | FALSE      | FALSE    | FALSE      | FALSE    |
| Intellectual Disability (TRUE/FALSE, IQ if known) | TRUE, ≤25 | Unknown | FALSE, FSIQ = 116 | FALSE, FSIQ = 91 | TRUE – mild, FSIQ = 63 | FALSE | TRUE, FSIQ = 68 | FALSE | TRUE, ≤25 | FALSE, FSIQ = 91 |
| Language level (at age of evaluation) | Delayed, single words | Delayed, no words | Delayed, single words | Delayed, phrase speech (sentences) | Delayed, phrase speech (ADOS mod. 2) | Delayed, phrase speech (ADOS mod. 2) | Delayed, phrase speech (ADOS mod. 2) | Unknown | Unknown | Unknown | Unknown |
| Language regression (TRUE/FALSE) | FALSE     | FALSE   | Unknown | TRUE       | FALSE      | TRUE       | FALSE    | TRUE     | TRUE     | TRUE       | TRUE     | TRUE       | TRUE     |
| Co-morbid psychological diagnoses | Learning disorders, motor skills delay, speech articulation problems, feeding disorder | None | None | None | None | None | None | None | None | None | None | None | None |
| Early motor delay (TRUE/FALSE) | FALSE     | TRUE     | FALSE   | FALSE      | FALSE      | FALSE      | FALSE    | FALSE    | FALSE    | FALSE      | 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. MPC scores are listed for missense mutations. All phenotypic information for SPARK participants was collected online.
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 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)
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. 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. One of the newly statistically significant genes we identified, FEZF2, is a powerful master regulator gene critical for establishing corticospinal neurons, 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.

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. 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. 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 (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. 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. 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
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 (BRSK2), and neuronal migration of pyramidal neurons (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 65.58 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**

**Participants**

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,** Social Communication Questionnaire-Lifetime and Developmental Coordination Disorder Questionnaire score60 (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, and regression was determined from the Autism Diagnostic Interview-Revised. 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 QGD-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 Vcycle 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. Sequences were filtered to a minimum standard of >85% of target covered at 20x, and on average, 96% of the target was sequenced to 20x. 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, duplicates were marked using Picard version 1.93 MarkDuplicates, and indels were realigned using GATK version 2.5-2-gf57256b IndelRealigner. Quality checks were performed on the BAM files using SAMTools 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 >10x coverage.

KING 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 ch6 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 and GATK. FreeBayes version v1.10.3-g961e65f3 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 MappingQualityRankSumTest –A MappingQualityZero –A FisherStrand –A HaplotypeScore –A MappingQualityRankSumTest –A MappingQualityZero –A QualByDepth –A RMSMappingQuality –A ReadPosRankSumTest. We required the 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, FreeBayes v1.1.0-441, and Platytypus 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 <20 | FS >80 | MQ <40.0 | MQRankSum <-12.5 | ReadPosRankSum <-8.0. For indels, we filtered out: QD <20 | FS >2000.0 | 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. Platytypus: variant calling was performed with local assembly analysis when at most ten haplotypes were allowed. Variants were filtered out for allele bias (p-value <0.05), 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 DNFilter algorithm that was retrained with the SSC data set. 1.800 de novo mutations identified by both lossoffos and Krumm et al. 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 novo variants 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 HaplotypeCaller to generate individual level gVCF files. All samples in the cohort were then jointly genotyped and have variant quality recalibrated by GATK. A variant was present in the offspring but not in the reference 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 ≥0.50 (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 but only in GenCODE Basic v19. Candidate variants in the ACMG secondary findings v29 gene list except PTEN, TSC1, and TSC2 were included. Coding de novo variants—non-sense, 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, reference GRCh37/38, SFARI Gene scores (version q1, 2018, https://gene.sfari.org/database/gene-scoring/), CADD, and findings from Deciphering Developmental Disorders project (gene2phenotype).

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 homoygotes reference, >0.9 for homozygotes alternative, and 0.3–0.7 for heterozygotes SNVs (0.25–0.75 for heterozygous 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 frameshift 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 v0.3.15 using default parameters in >60% of the GENCODE transcripts.

Identification of mosaic mutations. Mosaic SNVs were independently called in two centers—Oregon Health & Science University (OHSU) and CUMC. The OHSU approach was previously published 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 distribution and an FDR based approach to determine per site thresholds.

OHSU. SNVs were called as previously described. 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 × 10⁻⁵ -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_msk, 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. (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 (v 1.3.1-42) and BCFTools (v 1.3.1-174). 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. 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. CADO, MHC 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 allele 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 ≤ 1 × 10⁻⁶) across all populations represented in gnomAD and ExAC databases. To account for variations in the reference genome that are more challenging to resolve, we removed variant sites found in regions of nonuniqueness mappability (score < 1; 300 bp), likely segmental duplication (score > 0.95), and known low complexity. We found in regions of nonunique mappability (score < 1; 300 bp), likely gnomAD and ExAC databases). To account for regions in the reference estimation. 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 probability that a mosaic variant is actually a de novo 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. Call sets 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. Since confidence CNVs were also included for 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, brain critical genes, ASD significant genes, and ASD related genes. They have their coding regions overlapping with the CNV. CNVs greater than or equal to 50 kb in size were annotated with morbidty map case and control frequencies using a 50% reciprocal overlap while CNVs < 50 kb were annotated with their frequency in the 1000 genomes project 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 variant was annotated with pLI (ExAC release 0.3, http://exac.broadinstitute.org/downloads), ASD, RVS, LGD, and SFARI Gene scores (version q1, 2018, https://gene.sfari.org/database/gene-scoring). dncCNVs that affect DXS22 and any other candidate genes due to high variability in number of those regions among individuals.

UW, detection using XHMM and ConIFNER. CNVs from WES were called using ConIFNER and XHMM. ConIFNER version v0.2.2 was used with the $S$ value, –svd 7, set as a threshold as suggested by the scree plot. XHMM version 1.20 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/zeeeev/mergeSVcallers with the following parameters: -t xhmm,conifier -r 0.5 -s 50000; then merged VCF was sorted by Picard version v2.50, and zipped and indexed using Tabix version v0.2.6. Data was re-analyzed with 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 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 it had at least overall 50% reciprocal overlap. We re-genotyped each XHMM and CLAMMS CNVs identifi ed by one tool is covered by a set of CNVs predicted by the other tool. Predicted and validated CNVs were then compared with CLAMMS plots were manually investigated for each CNV in the final SF list. In addition, SF predictions were compared with PennCNV calls from array data, which had a confidence score of at least 10. 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 version 1.4 on a per chromosome basis normalizing by the relationship of the number 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. Briefly, the trinucleotide sequence 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).

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. 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. 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 Gamma (6,1) and Gamma (6,1). The prior mean relative risks were estimated using the relationship between burden and relative risk as described previously. 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 combining 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 following the method of Parikshak et al. The expression scores were measured by a r-statistic comparing the expression level in each layer against all other layers. Two candidate ASD risk genes (PAXL, KCNQ2) were not included in the layer enrichment analysis. Two candidate ASD risk genes (PAXL, KCNQ2) were not included in the layer enrichment analysis. 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). Two candidate ASD risk genes (PAXL, KCNQ2) were not included in the layer enrichment analysis. The mouse orthologs of human genes were retrieved from the FlyBase database. 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. To analyze the overall trend of specificity of a gene set, the mean specificity measure of its genes was compared with
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 \(^\ddagger\) 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 \(^\ddagger\ddagger\) 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. \(^1\ddagger\)

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 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 \(^4\) 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 \(^1\) 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/.
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 Pediatr.* 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. Tamminen, 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 Rubbies, 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 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* 516, 464–472 (2014).

13. Jacquemont, S. et al. A higher mutational burden in females supports a female protective mechanism 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. Defining 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, *Neurol. Engl. J. Med.* 338, 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. Han, W. et al. The contribution of mosaic variants to autism spectrum disorder. *JAMA Pediatr.* 173, 2231–2234 (2017).

27. Levy, J. et al. N4RA2 haploinsufficiency is associated with intellectual disability and autism spectrum disorder. *Clin. Genet.* 94, 264–268 (2018).

28. Epi, K. C. et al. De novo mutations in epileptic encephalopathies. *Nature* 501, 217–221 (2013).

29. Guo, H. et al. Inherited and multiple de novo mutations in autism risk genes suggest a multifactorial model. *Mol. Autism* 9, 64 (2018).

30. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466, 368–372 (2010).

31. Maussion, G. et al. Implication of LRRC44 and DPPE in neurodevelopmental disorders. *Am. J. Med. Genet.* A 173, 395–406 (2017).

32. Parikshak, N. N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 997–1007 (2013).

33. Parkishak, N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* 155, 1008–1021 (2013).

34. Lodato, S. et al. Gene co-regulation by Fzd2 selects neurotransmitter identity and connectivity of corticospinal neurons. *Science* 357, 404–407 (2017).

35. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* 508, 199–206 (2014).

36. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* 50, 825–833 (2018).

37. Han, W. et al. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.* 9, e1003671 (2013).

38. Lodato, S. et al. Gene co-regulation by Fzd2 selects neurotransmitter identity and connectivity of corticospinal neurons. *Science* 357, 404–407 (2017).

39. Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* 508, 199–206 (2014).

40. Brueggeman, L., Koomar, T. & Michaelson, J. Forecasting autism gene discovery efficiency is associated with intellectual disability and autism spectrum disorder. *J. Autism Dev. Disord.* 32, 2231–2234 (2012).

41. Williams, S. et al. The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.* 12, https://doi.org/10.1101/148353 (2017).

42. Roak, B. J. et al. Recurrent de novo mutations implicate novel genes underlying autism. *Nat. Neurosci.* 20, 747–754 (2015).

43. 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).

44. Yuen, R. K. et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* 21, 185–191 (2015).

45. 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 (2011).

46. Kishi, M., Pan, Y. A., Crump, J. G. & Sanes, J. R. Mammalian SAD kinases are required for neuronal polarization. *Science* 307, 929–932 (2005).

47. Johnson, P. 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).

48. Sandin, S. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. (2019). https://www.bionx.org/content/10.1101/484113v3.

49. 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).

50. Rutter, M., Bailey, A. J. & Lord, C. Social communication questionnaire (SCQ). (2003).

51. Wilson, B. N. et al. Psychometric properties of the revised Developmental Coordination Disorder Questionnaire. *Phys. Occup. Ther. Ped.* 29, 182–202 (2009).

52. 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).

53. 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).
