nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0107-y

An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder

Donna M. Werling[1,28], Harrison Brand[2,3,4,28], Joon-Yong An[1,28], Matthew R. Stone[2,28], Lingxue Zhu[5,28], Joseph T. Glessner[2,3,4], Ryan L. Collins[2,3,6], Shan Dong[1], Ryan M. Layer[7,8], Eirene Markenscoff-Papadimitriou[1], Andrew Farrell[7,8], Grace B. Schwartz[1], Harold Z. Wang[2], Benjamin B. Currall[2,3,4], Xuefang Zhao[2,3,4], Jeanselle Dea[1], Clif Duhn[1], Carolyn A. Erdman[1], Michael C. Gilson[1], Rachita Yadav[2,3,4], Robert E. Handsaker[4,9], Seva Kashin[4,9], Lambertus Klei[10], Jeffrey D. Mandell[1], Tomasz J. Nowakowski[1,11,12], Yuwen Liu[13], Sirisha Pochareddy[14], Louw Smith[1], Michael F. Walker[1], Matthew J. Waterman[15], Xin He[13], Arnold R. Kriegstein[16], John L. Rubenstein[1], Nenad Sestan[14], Steven A. McCarroll[4,9], Benjamin M. Neale[4,17,18], Hilary Coon[19,20], A. Jeremy Willsey[1,21], Joseph D. Buxbaum[22,23,24,25], Mark J. Daly[4,17,18], Matthew W. State[1], Aaron R. Quinlan[7,8,20], Gabor T. Marth[7,8], Kathryn Roeder[5,26], Bernie Devlin[10]*, Michael E. Talkowski[2,3,4,27]* and Stephan J. Sanders[1]*

Genomic association studies of common or rare protein-coding variation have established robust statistical approaches to account for multiple testing. Here we present a comparable framework to evaluate rare and de novo noncoding single-nucleotide variants, insertion/deletions, and all classes of structural variation from whole-genome sequencing (WGS). Integrating genomic annotations at the level of nucleotides, genes, and regulatory regions, we define 51,801 annotation categories. Analyses of 519 autism spectrum disorder families did not identify association with any categories after correction for 4,123 effective tests. Without appropriate correction, biologically plausible associations are observed in both cases and controls. Despite excluding previously identified gene-disrupting mutations, coding regions still exhibited the strongest associations. Thus, in autism, the contribution of de novo noncoding variation is probably modest in comparison to that of de novo coding variants. Robust results from future WGS studies will require large cohorts and comprehensive analytical strategies that consider the substantial multiple-testing burden.

The rapid progression of genomics technologies, coupled with expanding cohort sizes, has led to significant progress in characterizing the genetic architecture of complex disorders[1–6]. Thus far, studies have mainly focused on genotyping array technologies to survey common variants and large, rare copy number variations (CNVs), as well as whole-exome sequencing (WES) to scan rare protein-coding variants. Common variant genome-wide association studies (GWAS) have been particularly successful in adult-onset disorders, and most loci discovered are in the noncoding genome[7]. In early-onset disorders with reduced fecundity, including autism spectrum disorder (ASD)[8], discovery has been largely driven by the identification of extremely rare, gene-disrupting de novo mutations that confer considerable risk[4,5,9].

WGS offers the opportunity to assay the contribution of rare variation in the noncoding genome, a potentially large and hitherto unexplored class of variation. Because noncoding variants mediate the specificity of gene expression at particular developmental stages and in particular tissues and cell types, identifying such variants could provide important insights into the biology underlying complex disorders[10–12]. However, interpreting WGS in the noncoding genome presents considerable challenges[13]. There are no reliable estimates of the number of loci that could mediate risk, the extent of such risk, nor the genomic characteristics of such loci—keys to predicting the success of this type of endeavor. Moreover, a noncoding equivalent is lacking for the triplet code in protein-coding regions[14], which has been critical for predicting which coding nucleotides will alter gene function when mutated. Any serious exploration of rare variants in the noncoding genome must acknowledge this uncertainty and account for the inevitable multiple comparisons that result, because failure to do so virtually ensures the detection of false-positive associations and erroneous biological conclusions. Therefore, WGS association studies will require the same unbiased approaches and statistical rigor that have been applied to linkage-, GWAS-, or WES-based gene discovery.

Here we present such an analytical framework and apply it to a family-based cohort. These analyses focus on ASD families with both affected and unaffected children because of the well-documented contribution of de novo mutations to ASD and the existing genomic data that allow us to target families without known genetic risk factors[4]. Specifically, we examine 519 ASD cases, their unaffected sibling

*A full list of authors and affiliations appears at the end of the paper.

controls, and both of their parents (2,076 individuals; Supplementary Table 1) from the Simons Simplex Collection (SSC)[15]. De novo mutations were annotated at the level of nucleotides, genes, and regulatory regions to define 51,801 annotation categories. In a category-wide association study (CWAS), no annotation category achieved statistical significance; furthermore, many biologically plausible noncoding categories that were enriched in controls achieved equivalent levels of significance as those that were enriched in cases. We did not observe evidence of a noncoding category comparable to de novo loss-of-function coding mutations in terms of both effect size and frequency. We have made this analytical framework publicly available, along with the necessary annotation data.

Results

Cohort selection and characteristics. All 519 ASD cases were selected from the SSC on the basis of the absence of de novo loss-of-function mutations or large de novo CNVs in prior WES and microarray data, with the objective of enriching this sample for undiscovered de novo variation. The majority of cases (92%; $n = 479/519$) were selected randomly after this exclusion, while the remaining 8% were selected for a pilot study[16] and were enriched for factors associated with increased de novo burden: older fathers, female cases, and cases with nonverbal IQ ≤ 70 (Supplementary Table 1).

Identification of single-nucleotide variants and insertion-deletions. Single-nucleotide variants (SNVs) and small insertion-deletions of <50 bp (indels)[17] were discovered using the Genome Analysis Toolkit (GATK)[18]; family structure was leveraged to define high-quality calls (Supplementary Figs. 1 and 2). Overall, we identified 3.7 million autosomal variants per individual, including 3.4 million SNVs and 0.3 million indels. Six algorithms were employed to detect de novo SNVs and indels outside of low-complexity regions, and 1,638 previously validated de novo mutations (1,477 SNVs; 161 indels) were used to distinguish high-confidence mutations (Supplementary Figs. 1 and 2, and Supplementary Table 2). Using independent experimental validation, confirmation rates compared favorably with published literature for both SNVs (96.8%; 212/219) and indels (82.4%; 145/176) (Methods and Supplementary Table 3)[16]. Both WGS and WES data were available for 990 children. In GENCODE-defined autosomal coding regions, 1,116 de novo variants (1,075 SNVs, 41 indels) were detected by WGS as compared to 896 de novo variants (869 SNVs, 27 indels) detected by WES (Supplementary Table 4). Of the 896 de novo WES variants, 870 were detected in the WGS data (97%; 849 SNVs, 21 indels), and 768 of these met our quality criteria (88% of 870; 754 SNVs, 14 indels). WGS identified an additional 348 de novo mutations (321 SNVs, 27 indels), in large part due to limited coverage in the WES data (Supplementary Table 4), including 19 predicted to result in loss of function and 58 missense variants predicted to be probably damaging by PolyPhen-2. Considering variants not detected in exome analysis in all 1,038 children, we observed 24 de novo loss-of-function mutations, including 3 case mutations in genes predicted to be loss of function intolerant (*CLCN3*, *FNBP4*, *PHIP*), and 259 missense mutations, including 7 probably damaging case mutations in genes predicted to be loss of function intolerant (Supplementary Table 4).

In WGS data, we observed a median of 64 de novo SNVs and 5 de novo indels per child across the autosomes, for a total of 72,298 variants (66,366 SNVs, 5,932 indels) (Supplementary Table 5). We saw a slight excess of mutations in cases as compared to their sibling controls that remained after applying linear regression to adjust for quality metrics influencing de novo mutation detection (relative risk (RR) = 1.024, $P = 0.0006$ for all mutations; RR = 1.024, $P = 0.001$ for noncoding mutations alone; one-sided binomial test; Supplementary Fig. 3). However, when we corrected for the effect of paternal age, which is known to influence mutation rates[19,20], no significant difference in de novo burden remained (RR = 1.006, $P = 0.2$ for all variants; RR = 1.005, $P = 0.24$ for noncoding mutations alone; Supplementary Fig. 4). Correction for all covariates, including paternal age and sequencing quality, was applied to all subsequent tests of de novo mutation burden.

Case–control association tests of SNVs and indels. The sheer diversity and complexity of noncoding functional annotations necessitates a strategy to interpret the multiple parallel hypotheses they evoke. For gene-based analyses, we used GENCODE gene definitions and surveyed four coding categories (for example, missense) and seven noncoding categories (for example, UTRs) (Fig. 1a). In all analyses, we compared the number of de novo mutations that mapped to these regions in cases to the number in sibling controls and then assessed the significance of these comparisons using 10,000 within-sibship case–control label-swapping permutations. *P* values were calculated as the proportion of permutations with RR as or more extreme than in the observed data, taking into account the direction of the observed RR (case burden, RR > 1; control burden, RR < 1). For categories with empirical $P < 0.01$, we ran another 90,000 permutations for accuracy. This analytical approach is used throughout the manuscript unless otherwise noted. After correcting for multiple comparisons, no significant excess of de novo variants in any gene-defined category was observed (Fig. 1a). Repeating the analysis considering SNVs and indels separately and considering only variants within or near one of 179 genes associated with ASD (Supplementary Fig. 5) at a liberally-defined false-discovery rate (FDR < 0.3; Supplementary Table 6)[4], only an excess of predicted damaging de novo missense mutations was apparent, although both promoter regions and UTRs showed a trend toward enrichment in cases (Fig. 1b). Neither constrained genes[21] nor mRNA targets of fragile X mental retardation protein (FMRP)[22] yielded nominally significant categories. Similar results were obtained when considering only variants at nucleotides conserved across species.

We next extended our analyses to include additional subsets of noncoding variation and designed a CWAS to assess multiple hypotheses. We integrated five approaches to annotation: (i) gene sets implicated in ASD biology (for example, targets of FMRP); (ii) functional annotation (for example, chromatin state); (iii) conservation across species; (iv) type of variant (for example, SNV or indel); and (v) GENCODE gene definitions. In total, we surveyed 51,801 distinct combinations of annotation categories (Fig. 2 and Supplementary Table 7), comparing the burden of de novo mutations in cases versus controls for each category (Fig. 3a). Eschewing a priori hypotheses, we treated all tests equally. Illustrating the risks in testing a limited set of investigator-selected annotation categories without appropriate correction, we observed equivalent *P* values in the top categories enriched in either cases or controls, many of which would yield strong biological hypotheses (Table 1 and Supplementary Table 7). For example, the top category enriched among cases was from conserved indels near protein-coding genes within regions of weak transcription (chromatin state 5)[23], while the top category enriched in controls was noncoding SNVs near loss-of-function-intolerant genes within genic enhancers (chromatin state 6; Table 1).

Similar to the correlation structures seen in other forms of genome-wide analyses, many of these annotation categories encompass overlapping sets of variants and are thus dependent (Fig. 3b), raising the question of what constitutes an appropriate correction for multiple comparisons. To estimate this correction, we generated 20,000 simulated datasets of annotated mutations (Supplementary Note) and assessed the correlation of *P* values for the 51,801 categories across the simulations. Eigenvalue decomposition estimated 4,123 effective tests (Supplementary Fig. 6), leading to a category-wide significance threshold of 1.2×10^{-5} (Fig. 3a). No annotation category was within an order of magnitude of this threshold.

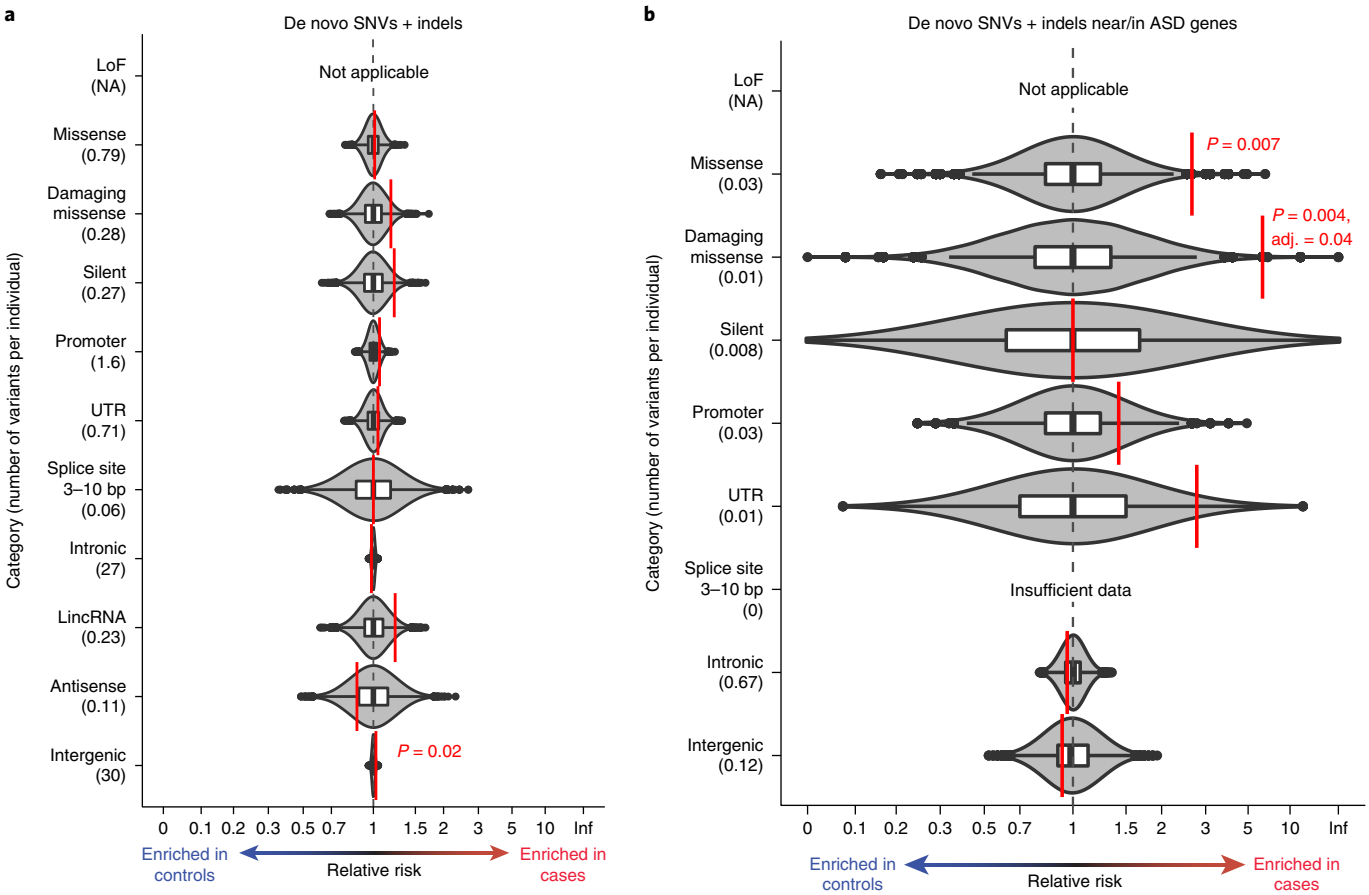

Fig. 1 | Burden analyses for gene-defined annotation categories. a, The observed relative risk of de novo mutations in cases versus controls is shown by the red line against gray violin plots representing the kernel density estimation of relative risk from 10,000 label-swapping permutations of case–control status for 11 gene-defined annotation categories. Box plots further illustrate the relative risk from permutations, including the median (center line), first and third quartiles (box), 1.5 times the interquartile range or the most extreme value (whiskers), and permuted relative risk observations beyond 1.5 times the interquartile range (outlier points). *P* values from a case–control label-swapping permutation analysis and Bonferroni-corrected *P* values (10 tests) ≤ 0.05 are shown. Loss-of-function (LoF) variants were not analyzed, as cases with such mutations were excluded from the cohort. **b**, The analysis in **a** repeated considering only de novo mutations in or near 179 ASD-associated genes. Permutation *P* values are Bonferroni corrected for seven tests. Considering SNVs and indels separately does not alter these findings (Supplementary Fig. 5).

k-means clustering of the simulated *P* values was used to identify the 200 most independent clusters of annotation categories (Fig. 3b–h). Testing the single category within each cluster with the most mutations per individual, a metric independent of cluster enrichment, also failed to identify a category within an order of magnitude of a significance threshold corrected for these 200 tests (2.5×10^{-4}; Supplementary Fig. 6).

We next considered whether there was evidence of a tendency toward enrichment of the 51,801 categories in cases, suggesting an underlying signal. We therefore counted the number of nominally significant categories and compared this to expectation (Fig. 3i–k). In coding regions, we observed more significant tests than expected in cases for SNVs and indels together ($P = 0.01$), but not in noncoding regions, either overall ($P = 0.20$) or near ASD-associated genes ($P = 0.64$). Notably, categories restricted to noncoding de novo indels showed a greater number of nominally significant results than expected ($P = 0.03$; Fig. 3h).

To explore the concept of an underlying signal further, we developed a polygenic risk score based on de novo variants, akin to similar scores developed previously for common and rare variants[24,25]. The rate of de novo mutations in cases and controls was weighted according to the category RR and adjusted for *P*-value correlation structure (Fig. 3b). Cross-validation was used to select annotation categories that best predicted case–control status. The resulting model included annotation categories relating to overall de novo burden (for example, all variants, all intergenic variants) and conservation scores across vertebrate species, but not coding regions or other functional annotations. The derived score accounted for only 0.31% of the variability in case status, which was not significantly different from zero.

Finally, we explored the impact of rare inherited SNVs and indels in the 405 families of European ancestry (Supplementary Table 1)[26]. Because runs of homozygosity (ROH) blocks often contain multiple homozygous variants inherited simultaneously, we counted only one variant per ROH block and excluded variants in ROH blocks that overlapped coding regions. No significant excess of rare homozygous (≤1% allele frequency) or heterozygous (≤0.1% allele frequency) SNVs and indels was observed overall or separately for maternally or paternally inherited variants, and no category reached significance in a CWAS (Supplementary Figs. 7 and 8).

Identification of structural variants. We next assessed whether structural variants (SVs), which rearrange large segments of the genome and can have important functional consequences, might demonstrate a noncoding signal. While much of the focus in SV detection from WGS has concentrated on CNVs alone, we previously

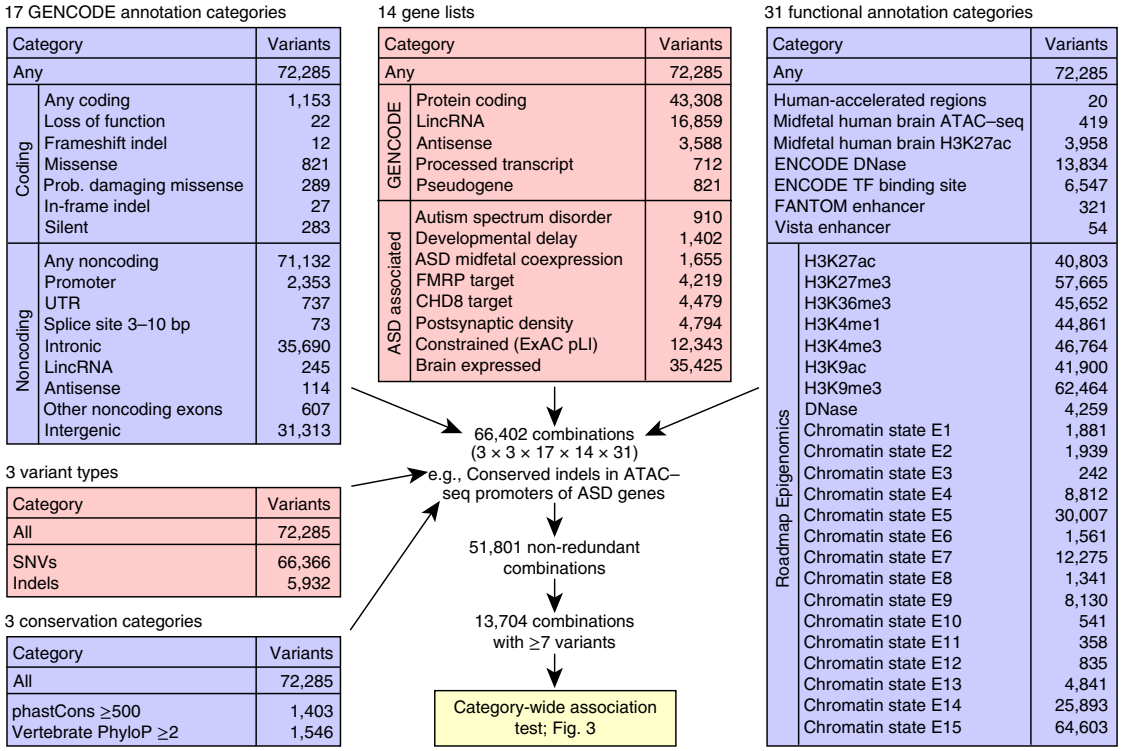

Fig. 2 | Defining annotation categories. Five groups of annotations were defined: (i) conservation across species; (ii) variant type; (iii) GENCODE gene definitions; (iv) gene lists; and (v) functional annotations. Picking one annotation from each group resulted in 66,402 possible combinations, of which 51,801 were non-redundant (Supplementary Table 7). The 13,704 annotation categories that included at least seven observed mutations were considered in the category-wide association test.

demonstrated the importance of translocations, inversions, and inversion-mediated complex SVs in ASD and congenital anomalies[27–29]. We thus characterized all classes of SV accessible to short-read WGS. Our SV discovery pipeline integrated eight algorithms to capture changes in read depth, clusters of reads with abnormal alignments, and mobile element insertions (Methods). We then developed an SV filtering pipeline to correct for the limited concordance among individual algorithms and several de novo prediction modules (Supplementary Figs. 9 and 10). Statistically significant CNV segments were integrated with predicted balanced SVs using a series of breakpoint-linking methods to identify signatures of 20 canonical, balanced, and complex SV classes (Supplementary Table 8)[27,30].

These analyses identified 98,790 SV sites and a median of 5,843 SVs per individual (Supplementary Fig. 11). These variants resulted in 101 likely loss-of-function and 30 whole-gene copy gain SVs per person; 4.1% of all SVs altered coding sequence as compared to 2.2% of SNVs and indels. We observed > 99% sensitivity and a 2.5% FDR for CNVs from WGS as compared to 1,087 CNVs previously reported from microarray data in these families (>40 kb; Supplementary Fig. 12)[4]. We relied on higher-resolution long-insert WGS (3.5-kb inserts, 102× median physical coverage) on 456 cases to validate SVs below microarray resolution (10–40 kb) and found a 5.2% overall FDR for 986 SVs (Supplementary Fig. 12).

We detected 171 de novo SVs from these 519 families, including 158 germline and 13 predicted somatic mosaic SVs. To facilitate reproducibility between studies, Supplementary Table 9 and Supplementary Data provide localization and visualization of each predicted de novo SV that can be evaluated by independent researchers. Validation assays could be designed for 168 de novo SVs using five complementary approaches, which revealed a 97.0% validation rate (163/168; Supplementary Table 9), including five subjects with sex chromosome aneuploidies (0.7% of cases and 0.2% of controls; Supplementary Fig. 13). We also observed 23 SVs initially predicted to arise de novo that demonstrated evidence of germline mosaicism in a parent (Supplementary Table 10). Collectively, this catalog of de novo SVs achieved high specificity, almost certainly at the cost of sensitivity for SV detection from short-read WGS, although there are few gold-standard datasets to estimate accurate SV mutation rates at present. Notably, a study from Turner et al. published during revision of this manuscript identified 88 de novo SVs from 476 of these quartets with an estimated 87.5% confirmation rate[31].

Association analyses of structural variants. Given the rarity of de novo SVs, there were limited data to derive insights comparable to de novo SNVs and indels (Fig. 3 and Supplementary Fig. 14). There was no significant difference in overall de novo SV burden between cases and controls (RR = 1.14, $P = 0.47$; Fig. 4a–c). There was a non-significant enrichment in cases for de novo loss-of-function SVs (1.3% in cases, 0.6% in controls; RR = 2.33; $P = 0.34$), which was more pronounced by removing mosaic SVs ($P = 0.07$) and included two SVs that disrupted ASD-associated genes: exonic deletion of *CHD2* (GRCh37.63:chr15:g.93484245_93488636del) and a balanced translocation of *GRIN2B* (t(12q21.2;13p11.2); Fig. 4d,e). The *GRIN2B* translocation emphasizes the importance of surveying all SV classes. Four other cases harbored SVs that disrupted constrained genes (pLI > 0.9)[21] that were not associated with ASD (*LNPEP*, *CYFIP1*, *SAE1*, *ZNF462*), while three occurred in siblings (*USP34*, *NUCKS1*, *STS*). No significant enrichments were detected in any class of noncoding variation or from CWAS-based analyses of SVs after correction for multiple testing (Supplementary Table 11), nor were significant associations detected from analyses of 19,643 rare inherited SVs (minor allele frequency (MAF) < 0.1%) and 441 rare homozygous deletions (MAF < 1%).

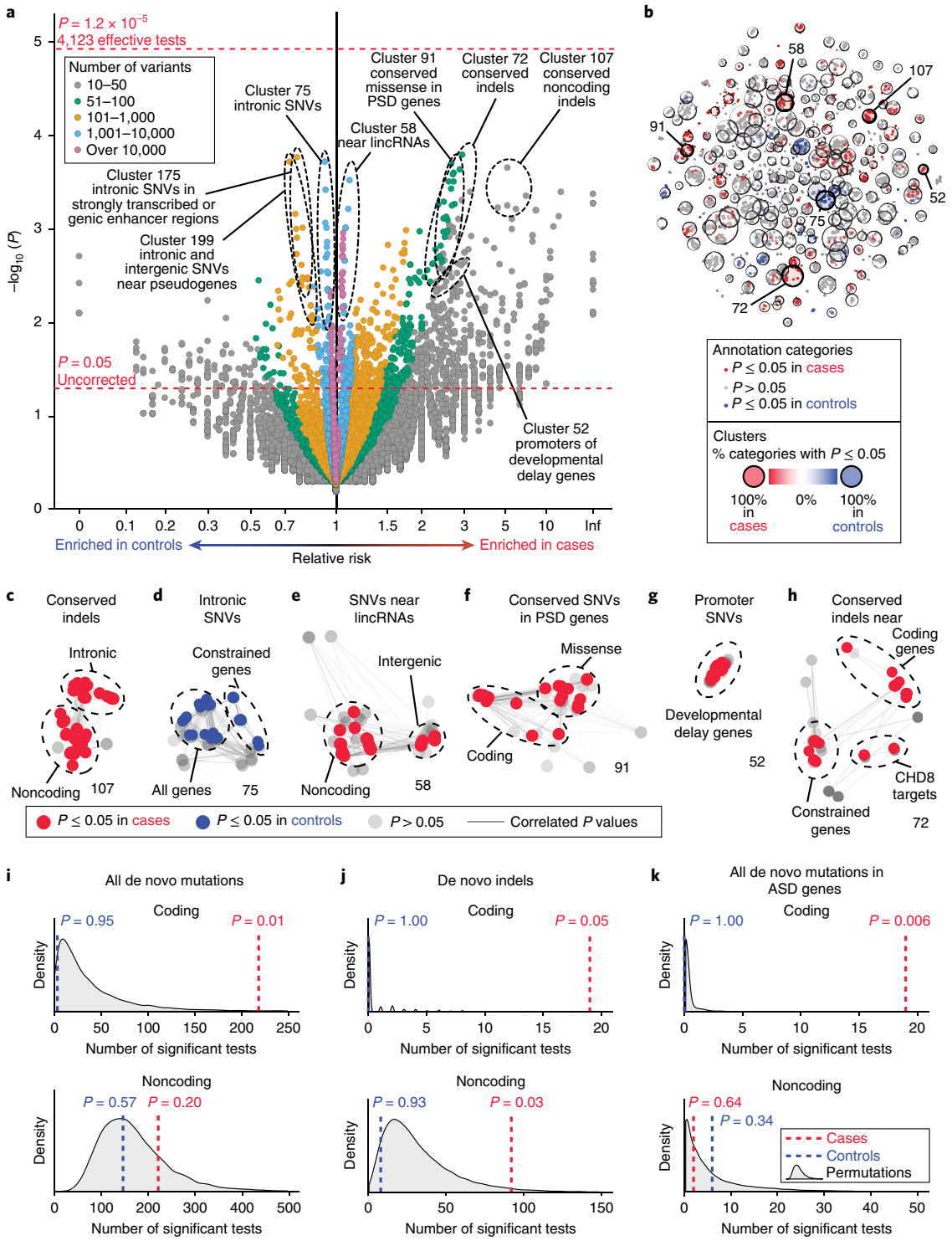

Fig. 3 | Category-wide association study. a, The burden of de novo SNVs and indels in 519 cases versus 519 controls for 13,704 annotation categories with ≥7 observed variants is shown as points in the volcano plot (Supplementary Table 7). Permutation *P* values were calculated by 10,000 label-swapping permutations of case–control status in each annotation category. No test survives Bonferroni correction for 4,123 effective tests (top horizontal red line). **b**, Correlations of *P* values between annotation categories (small dots) in simulated data are shown by proximity in the first two t-SNE dimensions. The large circles show 200 independent clusters of annotation categories defined by *k*-means clustering. Circle size represents the degrees of freedom accounted for by the cluster using Eigen value decomposition. In total, 4,123 effective tests explain 99% of the variability in *P* values (Supplementary Fig. 6). **c**–**h**, Six clusters from **b** are shown in greater detail, with cluster numbers in bold. The edges represent *P*-value correlation ≥0.4. **i**–**k**, The number of nominally significant annotation categories ($P \leq 0.05$ from two-sided binomial tests) was calculated for cases, controls, and 10,000 permutations to assess whether more annotation categories are enriched for de novo variants in cases than expected in **a**. Cases have a greater than expected number of nominally significant categories relating to coding mutations and noncoding indels, but not for all noncoding mutations nor for noncoding mutations nearest to ASD-associated genes. *P* values were calculated as the proportion of permutations in which the same or a greater number of categories had two-sided binomial test $P \leq 0.05$ as in the observed data.

Table 1 | Burden results for the most significant annotation categories from CWAS

Variant type	Most significant category within level of analysis	Variants per child (adjusted)	Relative risk	*P* value, uncorrected	Number of comparisons	*P* value, corrected
Cases: lowest *P* value per cluster in CWAS for top five clusters						
De novo indels	Conserved indels near protein-coding genes within chromatin state 5 (weak transcription) regions (cluster 72)	0.05	2.93	**0.0002**	4,123	0.66
De novo SNVs	Conserved coding SNVs within postsynaptic density genes (cluster 91)	0.06	2.63	**0.0002**	4,123	0.82
De novo indels	Conserved intronic indels within chromatin state 15 (quiescent) regions (cluster 107)	0.03	5.00	**0.0002**	4,123	0.91
De novo SNVs	Intergenic SNVs near lincRNAs underlying H3K36me3 (elongating) peaks (cluster 58)	4.69	1.11	**0.0003**	4,123	1.00
De novo SNVs and indels	Conserved coding variants in postsynaptic density genes within chromatin state 6 (genic enhancer) regions (cluster 23)	0.01	Inf	**0.0005**	4,123	1.00
Controls: lowest *P* value per cluster in CWAS for top five clusters						
De novo SNVs	Noncoding SNVs near constrained genes within chromatin state 6 (genic enhancer) regions (cluster 175)	0.47	1.36	**0.0002**	4,123	0.70
De novo SNVs	Intergenic SNVs near pseudogenes underlying H3K9me3 (constitutive repression) peaks (cluster 199)	0.41	1.44	**0.0002**	4,123	0.78
De novo SNVs	Intronic SNVs in constrained genes within chromatin state 5 (weak transcription) regions (cluster 75)	6.04	1.09	**0.0002**	4,123	0.78
De novo SNVs	Noncoding SNVs near constrained genes within chromatin state 5 (weak transcription) regions (cluster 20)	7.10	1.07	**0.001**	4,123	1.00
De novo SNVs and indels	Intergenic variants near genes coexpressed in midfetal brain within chromatin state 5 (weak transcription) regions (cluster 121)	0.08	1.86	**0.004**	4,123	1.00

De novo SNVs and indels from 519 cases and 519 controls were compared using a case–control label-swapping permutation analysis as described in the text. *P* values were Bonferroni corrected for 4,123 effective tests.

We observed nine cases (1.7%) and ten controls (1.9%) with large balanced chromosomal anomalies >3 Mb in length, as well as 35 CNVs > 40 kb not detected by microarray (Supplementary Fig. 12). Consistent with our previous analyses[27], rare SVs were more likely to cause genic loss of function than common SVs (odds ratio = 1.59; $P = 1.33 \times 10^{-30}$), particularly of constrained genes (odds ratio = 2.28; $P = 2.26 \times 10^{-9}$). However, there was no significant difference between cases and controls in overall SV size, percent of genome rearranged, or distribution of complex SVs (Fig. 4 and Supplementary Fig. 11). We also did not detect any changes in SV burden in proximity to genes or any signal when surveying up to 1 Mb from the transcription start site of all genes (minimum $P = 0.50$), constrained genes (minimum $P = 0.06$), or ASD-associated genes (minimum $P = 0.28$; Supplementary Fig. 15). Thus, despite dramatically improved access to the SV spectrum from WGS, we found no significant differences in the rate of rare inherited SV, nor did we observe evidence of significant biased transmission from either parent for any annotation category (Supplementary Table 12).

Power calculation. To estimate the required sample sizes, we performed a power calculation across estimates of RR and numbers of mutations per annotation category. Because of the complex correlation structure between categories, we used eigenvector analysis to estimate the effective number of tests conducted. This number increases with sample size, owing to increased likelihood of observing sufficient de novo mutations in any given annotation category to achieve significance: the number of effective tests increases from 4,123 at 519 families to ~7,600 at 4,000 families and approaches an asymptote of ~10,000 (Fig. 5 and Supplementary Fig. 6). The multiple-testing burden produces a threshold for statistical significance on the order of 5×10^{-6}. In this setting, over 8,000 families would be necessary to discover a noncoding element equivalent to missense variation. Further samples would likely be needed to hone in on a specific locus.

Discussion

We present an analytical framework for testing association between cases and controls in WGS data. Unlike for coding regions, there is no clear hypothesis of which noncoding regions harbor disease-causing rare variants in humans, nor is it understood which specific alleles are intolerant to mutation within those regions. To identify robust results, we have thus reasoned that WGS analyses must follow the same principles as other genomic analyses, acknowledging, defining, and correcting for the multiple comparisons that have been conducted, either explicitly or implicitly[32–35].

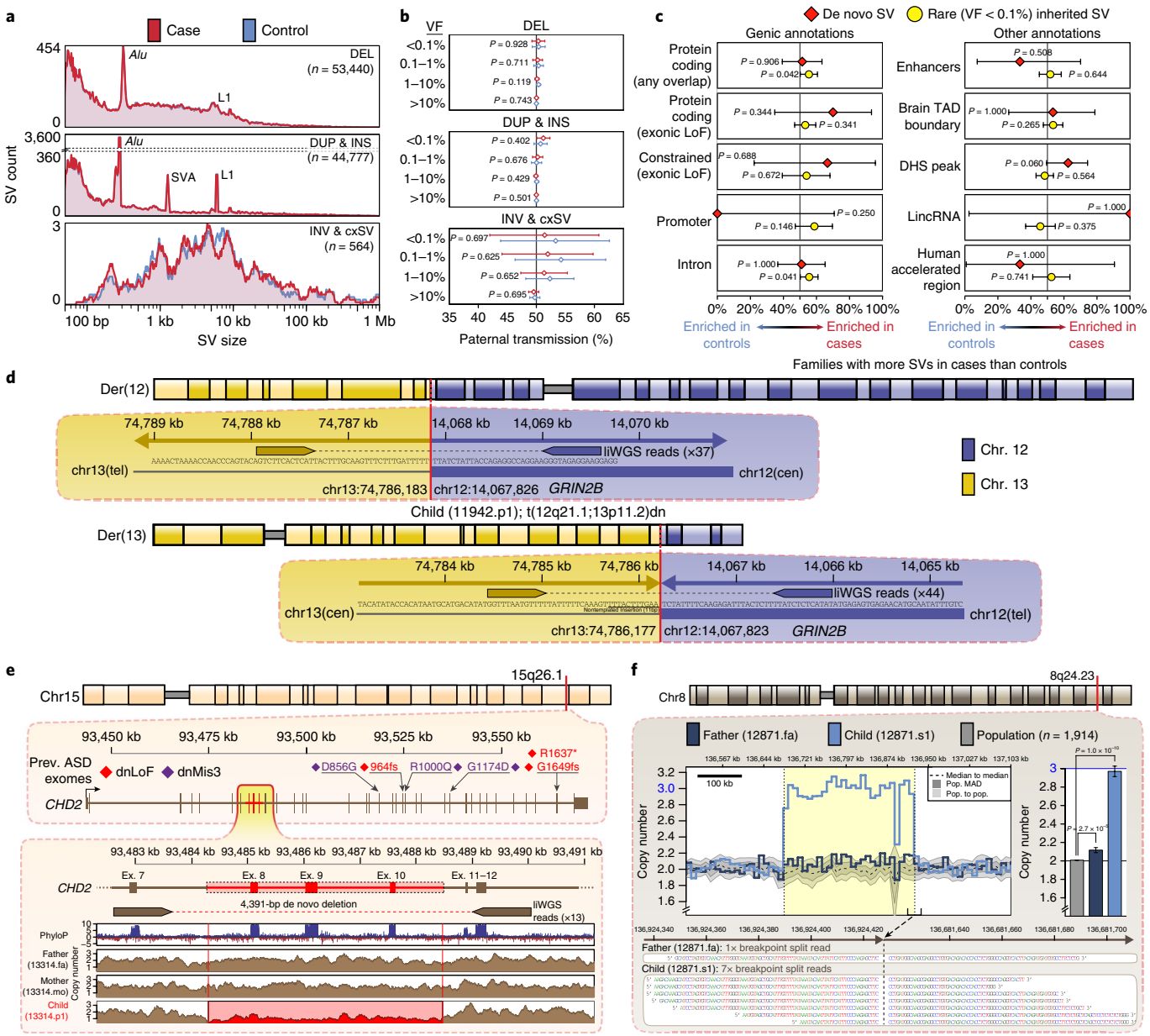

Fig. 4 | Structural variation in 519 ASD families. SV analyses identified a median of 5,843 SVs per genome, including 171 de novo SVs. **a**, We observed no difference in the distribution of SV sizes between cases ($n = 519$) and sibling controls ($n = 519$) for any class of SV (cxSV, complex SV) at an unadjusted nominal significance threshold (two-tailed Wilcoxon rank-sum test; $\alpha = 0.05$). **b**, We observed no differences in maternal/paternal transmission rates between cases and sibling controls for any class of SV or any range of variant frequencies (VF) (two-tailed binomial test). Mean paternal transmission rate (dot) and 95% binomial confidence intervals (error bars) are shown in the plot. **c**, We observed no significant enrichments for either de novo or rare inherited SVs (VF < 0.1%) in genic or noncoding annotations after correcting for multiple comparisons in a two-sided sign test between case and control counts. Error bars represent 95% confidence intervals. **d**, Analysis of balanced SVs discovered in a de novo reciprocal translocation in a case predicted to disrupt *GRIN2B* (t(12q21.2;13p11.2)), a constrained gene previously implicated in ASD[4,21]. **e**, WGS revealed small CNVs undetected by previous analyses, including a 4,391-bp de novo deletion of exons 8–10 of *CHD2* (GRCh37.63:chr15:g.93484245_93488636del), a gene previously implicated in ASD from de novo coding mutations[4]. **f**, Analysis of breakpoint sequences also classified 23 de novo SVs that were predicted to be germline mosaic in the parents, including a 242.8-kb paternally transmitted mosaic duplication at 8q24.23 that was previously characterized as de novo in the child (GRCh37.63:chr8:g.136681615_136924426dup). Bar plots represent the means and 95% confidence intervals of estimated copy number in the duplicated locus. All *P* values were calculated with a two-tailed *t* test of estimated copy numbers in sequential 36.4-kb bins.

By selecting annotations at the level of nucleotides, genes, and regulatory regions, we define 51,801 annotation categories and develop methods to test their association with ASD risk by de novo mutation burden. Considering the correlation structure of *P* values in simulated datasets, we determine the number of effective tests conducted, which increases with sample size but plateaus around 10,000 tests (Supplementary Fig. 6). These simulated data also allow us to define and test categories selected from independent annotation clusters, thereby permitting the use of a multiple-testing correction that depends only on the number of clusters and does not change with sample size (Supplementary Fig. 6). This CWAS approach is extendible to WGS association designs using different

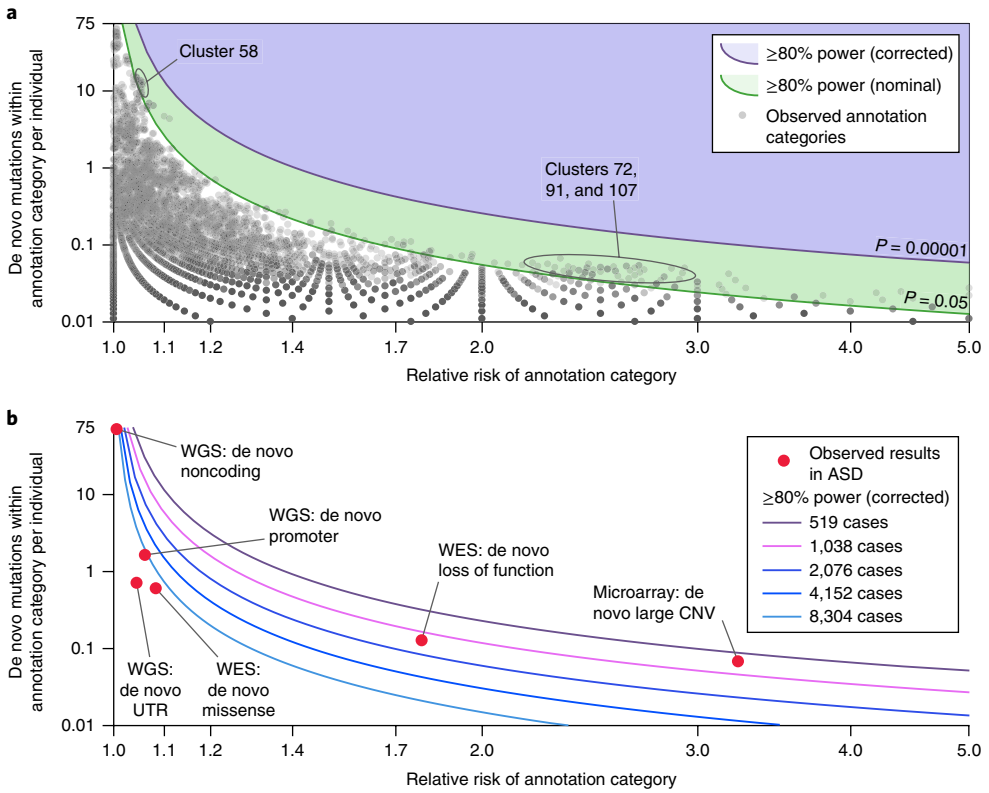

Fig. 5 | Effective number of tests in CWAS and power calculation. a, The green line shows the threshold to achieve 80% power at nominal significance across the range of relative risks of a category ($\log_{10}$-scaled *x* axis) and the number of de novo mutations per individual within the category ($\log_{10}$-scaled *y* axis). The purple line shows 80% power corrected for 4,123 effective tests. The gray dots represent the observed results for de novo mutation burden in 519 families for the 13,704 annotation categories with ≥7 mutations. **b**, The lines show the threshold of 80% power across the range of relative risks and category sizes as sample size increases (correcting for correspondingly more effective tests; Supplementary Note). For reference, the relative locations for six classes of variation are shown.

annotations. Our methods to accomplish this framework can be replicated using code hosted on Amazon Web Services on a publicly available Amazon Machine Image (for the most current AMI ID and SV pipelines, see https://github.com/sanderslab/WGS-pipeline and https://github.com/talkowski-lab/SV-Adjudicator, respectively).

Applying this framework to 519 families with a child affected by ASD, we are unable to demonstrate a rare noncoding variant contribution to ASD risk. Specifically, we do not observe association in 10 gene-defined categories (Fig. 1a), 200 independent annotation clusters (Supplementary Fig. 6), or 51,801 annotation categories (Fig. 3a). Furthermore, we do not observe an excess of nominally significant noncoding categories in cases (Fig. 3i) and could not develop an accurate predictor of case status using cross-validation. In contrast, the same techniques identified association of missense mutations in ASD genes (Fig. 1b), deleted exons in ASD genes (Fig. 4e), and an excess of nominally significant coding categories (Fig. 3i). Considering these results in the context of a power analysis (Fig. 5) gives important insight into genomic architecture, as it is unlikely that there is a class of noncoding variation equivalent to coding loss-of-function mutations in terms of both mutation frequency and effect size. These analyses also suggest that UTRs and promoter regions are likely to demonstrate association equivalent to, or weaker than, that of missense mutations. Finally, regulatory loci in intergenic and intronic regions, such as enhancers, are likely to be even harder to associate with ASD.

Prior to this analysis, this lack of power was predicted but not a foregone conclusion owing to the lack of equivalent systematic analyses of WGS data, the previous detection of ASD association in 225 families from WES[19,34,35], and previously reported nominally significant associations in ASD from WGS cohorts of fewer than 100 families[16,36]. Our estimates suggest that over 8,000 families would be required to demonstrate signal in a CWAS analysis such as that performed here (Fig. 5). Improved characterization of the noncoding functional genome, including through RNA-seq[37], ChIP–seq[38], Hi-C[39], and massively parallel reporter assays[40], could marginally bring this number down. Moreover, there are numerous WGS initiatives underway that will achieve such sample sizes in the near future[13] and necessitate an openly accessible, adaptable, and reproducible analysis framework to compare results across studies.

Our analyses provide some preliminary insights into the most likely noncoding risk factors that will emerge from larger samples. Our gene-defined analysis suggests that UTRs and promoters of ASD-associated genes could be the first categories to demonstrate noncoding risk (Fig. 1b). The CWAS analysis highlights the role of conserved indels, both in the 51,801 categories (Fig. 3a) and the 200 independent clusters (Supplementary Fig. 6). The burden analysis also identifies noncoding indels as a potential contributor (Fig. 3j), while the polygenic risk further implicates conservation across vertebrate species. Of note, by disrupting regulatory elements to a greater degree than SNVs while occurring far more often than SVs, indels could represent a sweet spot of statistical power for interrogating the noncoding genome.

Arguably, an alternative approach to WGS association designs might involve a priori prediction of which regulatory elements of the noncoding genome are important for disease risk, thereby limiting the number of tests evaluated and consequent statistical corrections. In terms of establishing a robust, unbiased framework to interpret disease association, we find this argument wanting. Perhaps the

simplest way to understand why is by analogy to candidate gene studies of complex disorders, which have had a miserable record regarding replication[41], with a plethora of false-positive results and a paucity of true-positive results[42]. This history should make us highly skeptical of methods based on investigator-selected a priori hypotheses in the noncoding genome. Continuing the analogy, instead of candidate genes, the field would be substituting 'candidate annotations' with all likelihood of poor outcomes, due to myriad combinations of annotations, cell types, brain regions, and developmental stages. Several ASD studies have selected different regions of the noncoding genome on which to focus[16,31,36,43,44], and associations from initial small studies have failed to replicate in larger datasets (Supplementary Fig. 15), a trend likely to persist if nominal significance is the threshold chosen for exploring genomes. The excess of missense mutations in postsynaptic density genes seen here is illustrative (cluster 91; Fig. 3a). We observe over 2.5-fold enrichment of these mutations in cases versus controls; however, analysis of exome data from 1,288 independent families[45] shows a much more modest 1.2-fold enrichment ($P=0.27$). This highlights 'winner's curse', in which the effect size in the discovery sample is likely to be greatly inflated[46], even for true associations.

Refinements in DNA sequencing, computing capability, and statistical analyses now permit simultaneous evaluation of the coding and noncoding genome and will eventually precipitate a sea change in how the impact of rare variation on disease risk is interpreted. Yet, the complexity of the noncoding genome complicates interpretation for both de novo and inherited variation, and there are perils in underestimating its complexity. Large-scale functional assays will continue to provide increasingly refined annotation of the regulatory genome, and perhaps eventually a noncoding equivalent to the triplet code will emerge. Until that time, we recommend the GWAS path for WGS studies: rigorous evaluation of multiple hypotheses and appropriate correction for that multiplicity, as we have outlined here. If we hold to these standards, very large sample sizes will be required to make headway, but we predict that the ensuing inferences will be sound and replicable.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0107-y.

Received: 19 July 2017; Accepted: 6 March 2018;
Published online: 26 April 2018

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
2. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
3. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
4. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
5. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
6. Marshall, C. R. et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat. Genet.* **49**, 27–35 (2017).
7. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45** (D1), D896–D901 (2017).
8. Power, R. A. et al. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry* **70**, 22–30 (2013).
9. Jin, S. C. et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat. Genet.* **49**, 1593–1601 (2017).
10. Visel, A. et al. ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
11. Shibata, M., Gulden, F. O. & Sestan, N. From trans to cis: transcriptional regulatory networks in neocortical development. *Trends Genet.* **31**, 77–87 (2015).
12. Silbereis, J. C., Pochareddy, S., Zhu, Y., Li, M. & Sestan, N. The cellular and molecular landscapes of the developing human central nervous system. *Neuron* **89**, 248–268 (2016).
13. Sanders, S. J. et al. Whole genome sequencing in psychiatric disorders: the WGSPD consortium. *Nat. Neurosci.* **20**, 1661–1668 (2017).
14. Caskey, C. T., Tompkins, R., Scolnick, E., Caryk, T. & Nirenberg, M. Sequential translation of trinucleotide codons for the initiation and termination of protein synthesis. *Science* **162**, 135–138 (1968).
15. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).
16. Turner, T. N. et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).
17. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
18. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
19. O'Roak, B. J. et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
20. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
21. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
22. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
23. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
24. Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016).
25. Purcell, S. M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185–190 (2014).
26. Chaste, P. et al. A genome-wide association study of autism using the Simons Simplex Collection: does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biol. Psychiatry* **77**, 775–784 (2015).
27. Collins, R. L. et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.* **18**, 36 (2017).
28. Talkowski, M. E. et al. Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell* **149**, 525–537 (2012).
29. Redin, C. et al. The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nat. Genet.* **49**, 36–45 (2017).
30. Brand, H. et al. Paired-duplication signatures mark cryptic inversions and other complex structural variation. *Am. J. Hum. Genet.* **97**, 170–176 (2015).
31. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
32. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
33. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
34. Neale, B. M. et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
35. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
36. Yuen, R. K. et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–191 (2015).
37. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
38. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
39. van Berkum, N. L. et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, e1869 (2010).
40. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
41. Johnson, E. C. et al. No evidence that schizophrenia candidate genes are more associated with schizophrenia than noncandidate genes. *Biol. Psychiatry* **82**, 702–708 (2017).

42. Farrell, M. S. et al. Evaluating historical candidate genes for schizophrenia. *Mol. Psychiatry* **20**, 555–562 (2015).
43. Munoz, A. et al. De novo indels within introns contribute to ASD incidence. Preprint at *bioRxiv* https://doi.org/10.1101/137471 (2017).
44. Brandler, W. M. et al. Paternally inherited noncoding structural variants contribute to autism. Preprint at *bioRxiv* https://doi.org/10.1101/102327 (2017).
45. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
46. Ioannidis, J. P. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).

Acknowledgements

We are grateful to the families participating in the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (SSC). This work was supported by grants from the Simons Foundation for Autism Research Initiative (SFARI 385110 to N.S., A.J.W., M.W.S., S.J.S.; 385027 to M.E.T., J.D.B., B.D., M.J.D., X.H., K.R.; 388196 to G.M., H.C., A.R.Q.; and 346042 to M.E.T.), the US National Institutes of Health (R37MH057881 and U01MH111658 to B.D. and K.R.; HD081256 and GM061354 to M.E.T.; U01MH105575 to M.W.S.; U01MH111662 to M.W.S. and S.J.S.; R01MH110928 and U01MH100239-03S1 to M.W.S., S.J.S., A.J.W.; U01MH111661 to J.D.B.; K99DE026824 to H.B.; U01MH100229 to M.J.D.), the Autism Science Foundation to D.M.W., and the March of Dimes to M.E.T. M.E.T. was also supported by the Desmond and Ann Heathwood MGH Research Scholars award. We thank the SSC principal investigators (A. L. Beaudet, R. Bernier, J. Constantino, E. H. Cook Jr, E. Fombonne, D. Geschwind, D. E. Grice, A. Klin, D. H. Ledbetter, C. Lord, C. L. Martin, D. M. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M. W. State, W. Stone, J. S. Sutcliffe, C. A. Walsh, and E. Wijsman) and the coordinators and staff at the SSC clinical sites; the SFARI staff, in particular N. Volfovsky; D. B. Goldstein for contributing to the experimental design; the Rutgers University Cell and DNA repository for accessing biomaterials; and the New York Genome Center for generating the WGS data.

Author contributions

Experimental design: D.M.W., H.B., J.-Y.A., M.R.S., J.T.G., M.J.W., X.H., N.S., B.M.N., H.C., A.J.W., J.D.B., M.J.D., M.W.S., A.R.Q., G.T.M., K.R., B.D., M.E.T., and S.J.S. Identification of de novo SNVs and indels: D.M.W., J.-Y.A., S.D., M.C.G., J.D.M., L.S., A.J.W., and S.J.S. Identification of structural variants: H.B., J.-Y.A., M.R.S., J.T.G., R.L.C., R.M.L., A.F., H.Z.W., X.Z., M.C.G., R.E.H., S.K., L.S., S.A.M., A.R.Q., G.T.M., and M.E.T. Confirmation of de novo variants: D.M.W., H.B., S.D., G.B.S., H.Z.W., B.B.C., J.D., C.D., C.A.E., R.Y., M.F.W., and M.J.W. Annotation of functional regions: D.M.W., J.-Y.A., S.D., E.M.-P., J.D.M., Y.L., S.P., J.L.R., N.S., M.E.T., and S.J.S. Generation of midfetal H3K27ac and ATAC–seq data: E.M.-P., T.J.N., A.R.K., and J.L.R. Development of genomic prediction score and de novo score: L.Z., L.K., K.R., and B.D. Analysis of SNVs and indels (Figs. 1–3): D.M.W., J.-Y.A., and S.J.S. Analysis of structural variants (Fig. 4): H.B., M.R.S., J.T.G., X.Z., and M.E.T. Assessment of *P*-value correlations, effective number of tests, and power analysis (Figs. 3 and 5): D.M.W., J.-Y.A., L.Z., G.B.S., K.R., B.D., and S.J.S. Manuscript preparation: D.M.W., H.B., J.-Y.A., M.R.S., L.Z., J.T.G., R.L.C., S.D., B.M.N., H.C., J.D.B., M.J.D., M.W.S., A.R.Q., G.T.M., K.R., B.D., M.E.T., and S.J.S.

Competing interests

J.L.R. is cofounder, stockholder, and currently on the scientific board of Neurona, a company studying the potential therapeutic use of interneuron transplantation. B.M.N. is an SAB member of Deep Genomics and serves as a consultant for Avanir Therapeutics. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0107-y.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.D. or M.E.T. or S.J.S.

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

[1]Department of Psychiatry, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [2]Center for Genomic Medicine and Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Neurology, Harvard Medical School, Boston, MA, USA. [4]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA, USA. [5]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Program in Bioinformatics and Integrative Genomics, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [7]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT, USA. [8]USTAR Center for Genetic Discovery, University of Utah School of Medicine, Salt Lake City, UT, USA. [9]Department of Genetics, Harvard Medical School, Boston, MA, USA. [10]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [11]Department of Anatomy, University of California, San Francisco, San Francisco, CA, USA. [12]Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA, USA. [13]Department of Human Genetics, University of Chicago, Chicago, IL, USA. [14]Department of Neuroscience and Kavli Institute for Neuroscience, Yale School of Medicine, New Haven, CT, USA. [15]Department of Biology, Eastern Nazarene College, Quincy, MA, USA. [16]Department of Neurology, University of California, San Francisco, San Francisco, CA, USA. [17]Analytical and Translational Genetics Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [18]Department of Medicine, Harvard Medical School, Boston, MA, USA. [19]Department of Psychiatry, University of Utah School of Medicine, Salt Lake City, UT, USA. [20]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT, USA. [21]Institute for Neurodegenerative Diseases, UCSF Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. [22]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [23]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [24]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [25]Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [26]Department of Computational Biology, Carnegie Mellon University, Pittsburgh, PA, USA. [27]Departments of Pathology and Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [28]These authors contributed equally: Donna M. Werling, Harrison Brand, Joon-Yong An, Matthew R. Stone, Lingxue Zhu.
*e-mail: devlinbj@upmc.edu; talkowski@chgr.mgh.harvard.edu; stephan.sanders@ucsf.edu

Methods

Sample selection. 519 quartet families (2,076 samples) with no known de novo rare CNVs, de novo loss-of-function mutations, or inherited rare CNVs at known ASD-associated loci in the proband were selected from the SSC (Supplementary Table 1). The first 40 families were additionally selected for high paternal age, low IQ, and female sex, while the second 479 were selected at random. All families had pre-existing microarray data[4] and WES (472 quartets and 47 proband trios)[45]. All subjects were consented for participation in genomic studies, data were deidentified by SFARI before sharing with researchers, and data access and analyses were approved by the UCSF IRB, Partners Healthcare IRB, and the University of Utah IRB.

Whole-genome sequencing. Whole-blood-derived DNA from all individuals was transferred from the Rutgers University Cell and DNA Repository (RUCDR) to the New York Genome Center (NYGC). Twenty-one families were excluded for low DNA quality, and the remaining 519 families were submitted for WGS. Data for the first 40 families were generated by PCR-based library preparation and Illumina HiSeq 2000, and PCR-free library preparation and Illumina HiSeq X Ten were used for the remaining 479 families. All sequencing used 150-bp paired-end cycles with a median insert of 423 bp. These data had a 99.3% median alignment rate, 0.50 strand balance, 0.11% duplication rate, and 37.8× median coverage per individual.

Data processing. Using the NYGC processing pipeline, FASTQ reads were aligned to the GRCh37.63 reference using BWA-mem v0.7.8-r455. Reads were sorted and duplicates were removed with Picard version 1.83. Indel realignment, base quality score recalibration, and variant calling were performed using GATK haplotype caller (GATK v3.1-1-g07a4bf8 for 19 batch 1 families, v3.2-2-gec30ce for 21 batch 1 families, and v3.4-0-g7e26428 for all 479 batch 2 families.

The BAM and gVCF files for all 2,076 samples were transferred to the Amazon Web Services (AWS) S3 storage system where they are accessible with approval from the Simons Foundation Autism Research Initiative (SFARI Base; https://sfari.org/resources/sfari-base). For downstream steps on AWS, we deployed CfnCluster on the Lustre cluster system using multiple m4.10xlarge instances. Using the GATK best-practices protocol, (https://software.broadinstitute.org/gatk/best-practices/; GATK v3.4-46-gbc02625), we merged individual gVCF files into a combined VCF and ran SNP and indel recalibration. Variant Quality Score Recalibration (VQSR) metrics were created from a training set of validated resources: dbSNP build 138, HapMap 3.3, 1000 Genomes OMNI 2.5, and 1000 Genomes Phase 1. For the following analysis, we excluded variant calls located in low-complexity regions[47] or with VQSR tranche 99.9–100%, as these calls have a high error rate or unusual characteristics[47,48]. Indels were realigned using left-normalization, and multiallelic variants were split into individual VCF lines using BCFtools[49].

Detection of high-quality SNVs and indels. As we had no established best practices for filtering rare variants in WGS data, we developed an optimized set of quality metric thresholds to detect rare SNVs and indels. For this, we compared two sets of rare variants with distinct quality metrics: (i) private transmitted variants observed in one family with no frequency information in 1000 Genomes or ExAC (likely true variants) and (ii) Mendelian violations in at least one child but also observed in an unrelated individual (likely false positives). We used receiver operating characteristic (ROC) curves to assess the ability of individual quality metrics to distinguish these true and false calls (Supplementary Figs. 1 and 2, and Supplementary Table 2). The metric and threshold that yielded the maximum increase in specificity and the minimum decrease in sensitivity were selected and applied as a filter to the training set. We repeated this process until we no longer observed improvement in sensitivity and specificity (details in the Supplementary Note).

Detection of high-quality de novo SNVs and indels. De novo SNVs were detected by four algorithms run on the default settings: TrioDeNovo[50], DenovoGear[51], PlinkSeq (https://atgu.mgh.harvard.edu/plinkseq/), and DenovoFlow. For de novo indels, DenovoGear was replaced with Scalpel[52]. DenovoFlow is a custom script that parses all possible Mendelian violations from each family, given GATK quality metrics. The union of these four algorithms predicted 86,921 Mendelian violation SNVs and 5,726 indels per child.

These numbers are large, suggesting a high false-positive rate. To identify high-quality de novo variants, we applied the same sequential ROC approach as above with true-positive calls defined by PCR Sanger validation of de novo mutations from prior work (1,302 selected SNVs, 95 selected indels) and with all variant- and individual-level quality metrics for the child and both parents (Supplementary Figs. 1 and 2). Using three additional metrics for SNVs, this analysis predicted 87.3% sensitivity and 98.8% specificity; using four additional metrics for indels, this analysis predicted 86.3% sensitivity and 93.0% specificity (Supplementary Table 2).

Validation of high-quality de novo SNVs. From the 66,366 high-quality de novo SNVs, 250 mutations were selected at random, conditional on available DNA, for validation in the child and both parents using PCR amplification and high-throughput sequencing on an Illumina MiSeq. In all analyses, we sought to both validate the presence of the putative variant and confirm de novo status based on absence of the variant in both parents. By investigation of off-target coverage, we determined that ≥50× depth was required for highly accurate genotyping of variants, and we used this threshold for all validation experiments. From the initial 250 variants, 13 either failed PCR amplification or MiSeq coverage thresholds in the proband, and an additional 18 variants failed coverage in at least one parent. Among the remaining 219 variants with successful assays, 212 were successfully validated, and all validated variants were confirmed to have arisen de novo (212/219; 96.8% confirmation rate; Supplementary Table 3).

Validation of high-quality de novo indels. We performed indel validation in two stages. In the initial exploratory analyses, 250 noncoding indels (125 deletions, 125 insertions) were selected at random from 9,961 high-quality de novo indel predictions for validation using the same PCR and MiSeq approach. Variants larger than 50 bp were excluded from the analysis (16 variants). Among the remaining 234 variants in the exploratory study, 22 variants were filtered out owing to failed PCR or insufficient MiSeq coverage in the probands (14 variants) or a parent (8 variants) (<50× depth). The remaining 212 putative mutations were examined with VarDict[53]. Among these, 137 were validated in the proband and 6 were determined to be inherited, for an overall confirmation rate in the exploratory analyses of 61.8% (131/212 variants with sufficient coverage in parents and child; Supplementary Table 3).

On the basis of these exploratory analyses, de novo indel prediction was refined (Supplementary Note), identifying 5,932 mutations overall, and a second round of validation was performed on 200 randomly selected variants of <50 bp in length. From this final validation set, 176 were successfully assayed and achieved adequate coverage in the child and parents; 148/176 variants were validated (84.1%), although 3 were determined to be inherited, yielding a final confirmation rate of 145/176 (82.4%) for de novo indel predictions, a substantial improvement over the exploratory analyses (Supplementary Table 3).

Validation of mutations near ASD-associated genes. Four putative mutations near known ASD-associated genes were also all validated as de novo: one SNV in the promoter of *ADNP* (GRCh37.63:chr20:g.49548007A>G), two SNVs near *GABRB3* (GRCh37.63:chr15:g.26327365A>G, GRCh37.63:chr15:g.26327513C>T), and one indel in the promoter of *NRXN1* (GRCh37.63:chr2:g.51259258delG).

SNV and indel annotation and statistical burden analyses. Variants were annotated to five annotation groups (Fig. 2) using ANNOVAR[54] and Bamotate[4]:

1. Variant type. Each variant was first classified by type, including SNV, indel (<50 bp), or SV (≥50 bp; deletions, duplications, insertions, inversions, and complex events);

2. Gene-defined annotation. Gene definitions from GENCODE (wgEncodeGencodeCompV19)[55] were obtained from the UCSC table browser (https://genome.ucsc.edu/), and variants were annotated using Bamotate; where multiple annotations were possible, variants were assigned in the following priority: coding, intron, promoter, UTR, and intergenic. Promoters were defined as the region 1 kb upstream of a transcription start site. The nearest transcription start site was identified for intergenic variants;

3. Species conservation. Variants were annotated to two conservation metrics: phastCons 46-way scores and phyloP scores from a 46-way vertebrate comparison from the UCSC table browser[56,57];

4. Gene sets. Gene lists associated with ASD were selected (for example, postsynaptic density genes). ASD risk genes (FDR < 0.3) were obtained from Sanders et al.[4]. Genes coexpressed with ASD risk genes were defined as the union of the two coexpression modules identified by Willsey et al.[58] in human midfetal prefrontal and primary motor-somatosensory cortex and infant mediodorsal thalamic nucleus and cerebellar cortex. Genes associated with developmental delay were downloaded from the Development Disorder Genotype-Phenotype Database (https://decipher.sanger.ac.uk/ddd; September 2016)[5,59]. The 2,156 genes were filtered to (i) confirmed developmental disorder genes, (ii) predicted loss of function, and (iii) include the term "brain" in the organ specificity list. CHD8 target genes were defined as the union of lists from two ChIP–seq studies[60,61], and FMRP target genes were selected from Darnell et al.[22]. Human postsynaptic density (PSD) proteins were downloaded from the Genes2Cognition database (http://www.genes2cognition.org/)[62]. Constrained genes were defined as having a probability of loss-of-function intolerance (pLI) score ≥0.9 in the ExAC database[21]. Either the transcript in which the variant was located or the nearest transcription start site (intergenic variants) was cross-referenced to these gene lists. For all gene lists, see Supplementary Table 6;

5. Regulatory regions. BED files were obtained for multiple regulatory regions. Vista enhancers were downloaded from the UCSC Genome Browser[63] (vistaEnhancers), and predefined enhancers were obtained from the FANTOM 5 server (http://enhancer.binf.ku.dk/presets/)[64]. ENCODE-defined transcription factor binding and DNase-hypersensitive sites were downloaded from the UCSC Genome Browser (wgEncodeRegTfbsClusteredV2 and wgEncodeRegDnaseClusteredV3). Human accelerated regions (HARs) were obtained from Doan et al.[65].

We also used histone marks and chromatin state data from the NIH Roadmap Epigenomics Project[66]. We merged data from brain tissues (E067 angular gyrus, E068 anterior caudate, E069 cingulate gyrus, E070 germinal matrix, E071

hippocampus middle, E072 inferior temporal lobe, E073 mid-frontal lobe, E074 substantia nigra, E081 fetal brain male, E082 fetal brain female), neurospheres (E053 neurosphere cultured cells cortex derived, E054 neurosphere cultured cells ganglionic eminence derived), ES-derived neuronal cells (E007 H1-derived neuronal progenitor cultured cells, E009 H9-derived neuronal progenitor cultured cells, E010 H9-derived neuron cultured cells), and astrocytes (E125 NH-A astrocytes).

We also used datasets generated at UCSF from midfetal human prefrontal cortex tissue (15–22 gestational weeks), including ATAC–seq (open chromatin) and ChIP–seq for H3K27ac (putative enhancers). Peaks were called by MACS (H3K27ac ChIP–seq) and Homer (ATAC–seq). Identified peaks common to two or more individual samples (≥1-bp overlap) were used for annotation.

Specific to our SV analysis, we investigated topologically associating domain (TAD) boundaries identified in fetal lung fibroblasts (IMR90) and embryonic stem cells (ESCs)[67]. The union of TAD boundaries was compiled from IMR90 and ESCs, and overlapping IMR90 and ESC boundaries were collapsed with BEDTools[68] and converted to hg19 coordinates by UCSC liftOver[69].

Burden testing for de novo SNVs and indels is described in the main text. All annotation reference files and software for annotation and burden testing of SNVs and indels can be accessed and implemented via a publicly available customized Amazon Machine Image on AWS (see https://github.com/sanderslab/WGS-pipeline for current AMI ID).

Detection of high-quality de novo structural variants. In our SV detection pipeline (Supplementary Fig. 9), we initially maximized sensitivity by integrating four paired-end/split-read (PE/SR) algorithms, three read-depth (RD) algorithms, and a mobile element insertion (MEI) detection pipeline to discover candidate SVs. We then adjudicated each predicted variant with a joint analysis of the cohort that included a series of modules for de novo variant filtering and a statistical test for de novo status. Our pipeline incorporated PE and SR calls from Delly (v0.7.3)[70], Lumpy (v0.2.13)[71], Manta (v.0.29.6)[72], and WHAM-GRAPHENING (v1.7.0)[73], each run jointly on all four family members; read-depth calls generated by NYGC from GenomeSTRiP (v2.00.1696)[74], CNVnator (v0.3.2)[75], and cn.MOPS (v1.8.9)[76]; and MEI calls from Melt (v2.0.5)[77] (additional details in the Supplementary Note). To determine the likelihood of a true SV, we developed an iterative random-forest-based modeling technique by testing for statistically significant differences between samples with and without SVs across four classes of orthogonal evidence types: (i) discordant PE read pairs, (ii) clipped SR, (iii) RD, and (iv) B-allele frequency (BAF) (additional details in the Supplementary Note). We used a batch-specific framework to jointly adjudicate SVs (pilot $n=160$ and Phase 1 $n=1{,}916$) to correct for demonstrable RD differences between the datasets (PCR^+ and PCR free, respectively) and further split the samples by sex for SV on allosomes. Metrics computed in the Phase 1 PCR-free samples were used whenever available. Across all CNVs that passed, we genotyped homozygous deletions, defined as samples with median normalized RD < 0.08. We identified five samples with sex chromosome anomalies: three XXY Klinefelter syndrome and two XYY syndrome (Jacob's syndrome). These variants were first detected from our initial depth assessment and then replicated by an independent algorithm[78].

In addition to polymorphic and de novo CNVs, we assessed balanced and complex SVs in the SSC, as we have done previously in this cohort with large SVs[27]. We applied the algorithm integration pipeline for PE/SR calls described above to find candidate inversion and translocation breakpoints and resolved the variant structure at these loci by matching the ordering of breakpoints to complex SV signatures previously identified by Collins et al.[27]. We identified 22,840 observations of 258 inversion-associated CNVs between 300 bp and 5 kb that were not found with the CNV discovery pipeline, as they lacked canonical PE/SR evidence and were below RD-only algorithm resolution. In total, we identified 53,440 deletions, 20,782 duplications, 23,995 insertions, 197 inversions, 4 reciprocal translocations, 5 sex chromosome aneuploidies, and 367 complex SVs across eight classes (Supplementary Table 8).

Validation of de novo SVs. We performed extensive validation of all putative de novo SV predictions using combinations of microarray, PCR with Sanger sequencing, PCR and long-read MiSeq sequencing, microarray, long-insert whole-genome sequencing (liWGS), and digital droplet PCR (ddPCR). Assays were attempted for at least one validation method on all de novo predictions, and a subset of variants were confirmed by multiple methods (see the Supplementary Note for complete validation details). Overall, we successfully designed assays for 168 of 171 de novo SV predictions (2 variants were from individuals who lacked sufficient DNA for confirmation and assays could not be designed for another variant owing to repetitive sequences at the breakpoint). We observed an overall validation rate of 97.0% (163/168) across all SV classes. All validation assays are provided in the Supplementary Information, including the number of assays performed and split reads from confirmation experiments for each variant (Supplementary Table 9). In addition, a series of visualizations were generated for each de novo SV prediction for each variant to enable visual inspection (Supplementary Data).

Comparison of SVs to microarray and long-insert WGS. We compared the performance of short-insert WGS (siWGS) SV calls to rare CNVs detected from liWGS ('jumping') libraries on 456 of the 519 cases[27] and microarray data from all 2,076 samples with SVs[4]. We performed the following filtering to correct for differences in resolution and sample differences across platforms: (i) microarray size threshold > 40 kb; a liWGS size range of 10–40 kb was used; (ii) repeat masking: SV comparisons were retained if ≤30% of the variant region overlapped an annotated segmental duplication region, microsatellite, heterochromatin, or one of our defined multiallelic regions (Supplementary Table 13); and (iii) all variant frequencies <10%. These filters resulted in 1,399 siWGS CNVs in the array comparison (Supplementary Fig. 12) and 986 variants in the liWGS comparisons (additional details in the Supplementary Note). Overall, we observed a 2.5% FDR and 99.6% sensitivity for microarray data and a 5.2% FDR and 91.9% sensitivity for liWGS (Supplementary Fig. 12).

SV annotation and statistical burden analyses. Each SV with any predicted overlap with the canonical transcript of 20,156 protein-coding genes (GENCODE v19) was annotated as genic. Deletions were considered loss of function if they affected any coding sequence, duplications were considered loss of function if they affected an exon but did not extend outside the transcript boundary, and inversions were considered loss of function if one breakpoint localized to a coding exon or any genic space spanning the coding sequence (but not if the entire coding sequence was inverted). Duplications were considered to be 'copy gain' if they spanned the entirety of a transcript. Intronic variants were required to localize fully to an intron. All variants, including noncoding variants, were additionally annotated with any gene whose UTR or promoter region (<1 kb upstream of the transcription start site) it disrupted (see Supplementary Fig. 14 for all SV annotations). Statistical burden testing was also performed using a CWAS design, paralleling the SNV analyses described above. Notably, families were selected after screening for probands harboring large de novo CNVs detected by microarray and de novo coding mutations detected by WES, but families with siblings harboring comparable mutations were not excluded. These analyses can impact estimates of de novo SV association, so we excluded any family where the sibling met similar exclusionary criteria ($n=27$; Supplementary Table 1). Rare SV analyses were restricted to the 405 families with European ancestry described in the SNV analyses.

Estimation of the number of effective tests in CWAS. We generated 20,000 sets of 72,285 autosomal, simulated variants randomly allocated to cases and controls, with the same proportion of SNVs and indels as in the observed data. These '1×' datasets simulating 519 families were combined at random to yield simulations at 2×, 4×, 8×, and 16× (8,304 cases). As with the CWAS analysis, we annotated these simulated variants against all 51,801 distinct annotation categories and tested categories for case–control burden using a one-sided binomial test, excluding categories with ≤7 variants in >50% of the simulations. We used z scores converted from P values to estimate correlations between annotation categories. We employed Eigen decomposition to estimate the number of effective tests within the genome-wide category space and k-means clustering to identify 200 clusters of correlated annotation categories.

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

Code availability. Our methods for de novo SNV and indel annotation and statistical analyses in WGS data can be replicated using code hosted on Amazon Web Services on a publicly available Amazon Machine Image (for the most current AMI ID, see https://github.com/sanderslab/WGS-pipeline), and SV analysis pipelines can be found at GitHub (https://github.com/talkowski-lab/SV-Adjudicator).

Data availability. All sequencing and phenotype data are hosted by the Simons Foundation for Autism Research Initiative (SFARI) and are available for approved researchers at SFARIbase (https://base.sfari.org/, SFARI_SSC_WGS_1b).

References

47. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
48. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
49. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
50. Wei, Q. et al. A Bayesian framework for de novo mutation calling in parents–offspring trios. *Bioinformatics* **31**, 1375–1381 (2015).
51. Ramu, A. et al. DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* **10**, 985–987 (2013).
52. Narzisi, G. et al. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods* **11**, 1033–1036 (2014).
53. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).

54. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10**, 1556–1566 (2015).
55. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
56. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
57. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
58. Willsey, A. J. et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997–1007 (2013).
59. 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).
60. Cotney, J. et al. The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6**, 6404 (2015).
61. Sugathan, A. et al. CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc. Natl. Acad. Sci. USA* **111**, E4468–E4477 (2014).
62. Bayés, A. et al. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat. Neurosci.* **14**, 19–21 (2011).
63. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
64. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
65. Doan, R. N. et al. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell* **167**, 341–354 (2016).
66. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
67. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
68. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
69. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
70. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
71. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
72. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
73. Kronenberg, Z. N. et al. Wham: identifying structural variants of biological consequence. *PLoS Comput. Biol.* **11**, e1004572 (2015).
74. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
75. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
76. Klambauer, G. et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).
77. Gardner, E. J. et al. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017).
78. Pedersen, B. S., Collins, R. L., Talkowski, M. E. & Quinlan, A. R. Indexcov: fast coverage quality control for whole-genome sequencing. *Gigascience* **6**, 1–6 (2017).

nature research

Corresponding Author: Michael Talkowski

Date: 01-18-17

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   519 quartet families (2,076 samples) were selected from the Simons Simplex Collection (SSC). As such, this was one of the largest WGS studies performed in the field to date (page 23). Sample size was limited by the cost of sequencing, however we provide extensive power analyses in Fig 5 to estimate required sample sizes corresponding with detection of associated relative risk.

2. Data exclusions

   Describe any data exclusions.

   The families were selected on the basis of a proband meeting diagnostic criteria for autism spectrum disorder and an unaffected sibling as well as both parents. Samples were excluded if previous analyses in this cohort detected a de novo copy number variant from chromosomal microarray analyses or a de novo loss of function mutation from exome sequencing. (page 23).

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   Findings were consistent between both pilot (n=40) and confirmation cohort (n=479)

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   N/A. Groups determined by diagnosis.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   N/A. We were not blinded to group allocation during data analysis, as pre-existing genetic data published for this cohort make blinding neither possible nor practical. To alleviate any complications from non-blinded analyses, genetic data from all case and control samples were simultaneously run through identical, automated variant calling protocols, optimized using variant validation data.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Our methods for de novo SNV and indel annotation and statistical analyses in WGS data can be replicated using code hosted on Amazon Web Services on a publicly available Amazon Machine Image (for the most current AMI ID, see https://github.com/sanderslab/WGS-pipeline), and SV analysis pipelines can be found at GitHub (https://github.com/talkowski-lab/SV-Adjudicator). (p14)

Software:
- BWA-mem v0.7.8-r455 for mapping sequencing reads to the reference genome
- Picard v1.83 for sorting reads and removing duplicate reads
- Picard v1.140 to generate sequencing quality metrics from BAM files
- GATK haplotype caller (v3.1-1-g07a4bf8 for 19 batch 1 families, v3.2-2-gec30ce for 21 batch 1 families, and v3.4-0-g7e26428 for all 479 batch 2 families) for SNV and indel variant calling
- CfnCluster v1.3.1 to deploy and maintain high performance computing clusters on Amazon Web Services, for downstream variant calling and filtering steps
- GATK v3.4-46-gbc02625 for merging gVCF files and running SNP and indel recalibration (Variant Quality Score Recalibration, VQSR)
- BCFtools v1.3 for normalization of indel calls and splitting multi-allelic variants, and also for identifying regions of homozygosity
- TrioDeNovo v0.04 for identifying de novo SNVs and indels
- DenovoGear v0.5.4 for identifying de novo SNVs and indels
- PlinkSeq v0.10 for identifying de novo SNVs and indels
- DenovoFlow, custom script for identifying de novo SNVs and indels given GATK quality metrics
- Scalpel v0.5.3 for identifying de novo indels
- bam-readcount v0.7.4 for sensitive SNV calling from BAM files, https://github.com/genome/bam-readcount
- VarDict v701398e for sensitive indel calling from BAM files
- IGV (Integrative Genomics Viewer) v2.3.67 and v2.3.81 for visual evaluation of called variants
- RepeatMasker v4.0.5 for screening variant regions for repeats and low complexity DNA sequences
- Annovar version 2016Feb01 for variant annotation
- Bamotate, custom script for variant annotation, available for use on public AWS AMI, see https://github.com/sanderslab for most current AMI

ID
- Bowtie2 v2.2.2 for mapping ChIP-seq and ATAC-seq reads to the reference genome
- MACS v2.1.1.20160309 for calling ChIP-seq peaks
- Homer for calling ATAC-seq peaks
- bedtools v2.25.0 and v2.26.0, for comparing and merging genomic region boundaries for annotation references, and for investigating CNV intervals and breakpoints
- Delly v0.7.3 for detection of structural variants
- Lumpy v0.2.13 for detection of structural variants
- Manta v0.29.6 for detection of structural variants
- WHAM-GRAPHENING v1.7.0 for detection of structural variants
- GenomeSTRiP v2.00.1696 for detection of structural variants
- CNVnator v0.3.2 for detection of structural variants
- cnMOPS v1.8.9 for detection of structural variants
- Melt v2.0.5 for detection of mobile element insertions
- Blat v.35 for assessing split reads during structural variantion validation
- indexcov for confirming sex chromosome aneuploidies using sequencing coverage data
- RdTest, read depth verification algorithm for SV calls, see https://github.com/talkowski-lab/SV-Adjudicator
- R v3.3.0 for exploratory analyses, statistical testing, and plotting
- Identity v1.0 script for confirming sample identify using genotypes, http://genomic-identity.wikidot.com
- PLINK v1.9 to check samples' sex, Mendelian error rate, and sample relatedness (identity by descent)
- Python v3.6.2 for exploratory analyses, structural variant processing, statistical testing, and plotting; Python v2.7 for merging label-swapping permutation output files from SNV and indel burden testing
- Jupyter v4.1.0 for exploratory analyses and plotting
- Snakemake v3.13.3 for structural variant processing workflows
- Numpy v1.13.3 for numerical analyses
- Scipy v1.0.0 for statistical testing and sparse graph clustering
- Pandas v0.21.0 for data processing and analysis
- Pysam v0.13 for analysis and processing of BAM and VCF files
- Pybedtools v0.7.10 for analysis of genomic intervals
- Matplotlib v1.5.1 for plotting
- Seaborn v0.7.1 for plotting
- Cython v0.26.1 for numerical and SAM flag analyses
- AWS CLI v1.10.56
- FBAT 2.0.4 for running transmission disequilibrium tests (TDT) for inherited variants
- Ansible v2.2.1.0-0.3.rc3 to deploy multiple AWS instances for permutations and burden testing

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	All sequencing and phenotype data are hosted by the Simons Foundation for Autism Research Initiative (SFARI) and are available for approved researchers at SFARIbase (https://base.sfari.org, accession SFARI_SSC_WGS_1b). (page 24 and 34)

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A. No antibodies were used in this study.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A. No cell lines were used in this study.

b. Describe the method of cell line authentication used.

N/A. No cell lines were used in this study.

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A. No cell lines were used in this study.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A. No cell lines were used in this study.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A. No animals were involved in this study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

All 519 ASD cases were selected from the SSC based on the absence of de novo loss-of-function mutations or large de novo CNVs in prior exome sequencing and chromosomal microarray data, with the objective to enrich this sample for undiscovered de novo variation. The majority of cases (92%, N=479/519) were selected randomly after this exclusion, while the remaining 8% were selected for a pilot study and were enriched for factors associated with increased de novo burden: older fathers, female cases, and cases with nonverbal IQ ≤70. (pg 5). Parents and one unaffected sibling of each proband were also included in this work.