nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0420-0

Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk

Jian Zhou[1,2,3,9], Christopher Y. Park[3,4,9], Chandra L. Theesfeld[1,9], Aaron K. Wong[3], Yuan Yuan[4,5], Claudia Scheckel[4,6], John J. Fak[4], Julien Funk[3], Kevin Yao[3], Yoko Tajima[4], Alan Packer[7], Robert B. Darnell[4*] and Olga G. Troyanskaya[1,3,8*]

We address the challenge of detecting the contribution of noncoding mutations to disease with a deep-learning-based framework that predicts the specific regulatory effects and the deleterious impact of genetic variants. Applying this framework to 1,790 autism spectrum disorder (ASD) simplex families reveals a role in disease for noncoding mutations—ASD probands harbor both transcriptional- and post-transcriptional-regulation-disrupting de novo mutations of significantly higher functional impact than those in unaffected siblings. Further analysis suggests involvement of noncoding mutations in synaptic transmission and neuronal development and, taken together with previous studies, reveals a convergent genetic landscape of coding and noncoding mutations in ASD. We demonstrate that sequences carrying prioritized mutations identified in probands possess allele-specific regulatory activity, and we highlight a link between noncoding mutations and heterogeneity in the IQ of ASD probands. Our predictive genomics framework illuminates the role of noncoding mutations in ASD and prioritizes mutations with high impact for further study, and is broadly applicable to complex human diseases.

Great progress has been made in the past decade in understanding the genetics of ASD, establishing de novo mutations, including copy number variants (CNVs) and point mutations that likely disrupt protein-coding genes, as important causes of ASD[1,2]. However, when combined, all the known ASD-associated genes explain only a small fraction of new cases and it is estimated that, overall, de novo mutations in protein-coding genes (including CNVs) contribute to no more than 30% of simplex ASD cases[2,3]. The vast majority of identified de novo mutations are located within intronic and intergenic regions; however, little is known regarding their contribution to the genetic architecture of ASD or for any other complex disease.

A potential role for noncoding mutations in complex human diseases including ASD has long been speculated. Human regulatory regions show signs of negative selection[4], suggesting that mutations within these regions lead to deleterious effects. Studies of inherited common variants have also shown enriched disease association in noncoding regions[5]. Furthermore, noncoding mutations that affect gene expression have been found to cause Mendelian diseases[6] and to be enriched in cancer[7]. Expression dosage effects have also been suggested to underlie the link between CNVs and ASD[8]. Recently, parentally inherited structural noncoding variants have been linked to ASD[9]. Also, in a small cohort of ASD families, some trends with limited sets of mutations have been reported[10–12]. Likewise, despite the major role that RNA-binding proteins (RBPs) have in post-transcriptional regulation, little is known of the pathogenic effect of noncoding mutations affecting RBPs (other than the effect of mutations in canonical splice sites). Thus, noncoding mutations could be a cause of ASD, but no conclusive connection between regulatory de novo noncoding mutations (either transcriptional or post-transcriptional) and the etiology of ASD has been established.

Recent developments make it possible to perform large-scale studies that reliably identify de novo noncoding mutations at the whole-genome scale. The Simons Simplex Collection (SSC) of whole-genome sequencing (WGS) data for 1,790 families differs from many previous large-scale studies in its design, which includes matched unaffected siblings[3,13–16]. These provide critical background controls for detecting excess mutation burden in probands, as it is otherwise hard to distinguish excess levels of mutations that are relevant to disease from irrelevant biological and technical variation, such as differences in genetic background or artificial biases originating from sequencing, variant calling and filtering procedures.

However, even with study designs using matched control individuals, detecting the contribution of de novo noncoding mutations is still challenging and establishing the role of the vast noncoding space in the genetic basis of autism remains difficult. Two recent studies[17,18] have demonstrated that, even when considering a wide variety of possible functional annotation categories (for example, mutations in known regulatory sites, mutations at the location of known histone marks and mutations near ASD- or disease-relevant gene sets), no significant signal specific to noncoding mutations in ASD probands was observed, and that approach would require a very large cohort to detect signal[17]. This is consistent with the expectation that noncoding mutations, in contrast to loss-of-function (LoF) coding mutations, can vary highly in functional impact, with potentially only a small fraction of variants having strong effect sizes. Thus, the challenge is to move beyond simple mutation counts, which are susceptible to both statistical power challenges and confounding

[1]Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. [2]Graduate Program in Quantitative and Computational Biology, Princeton University, Princeton, NJ, USA. [3]Flatiron Institute, Simons Foundation, New York, NY, USA. [4]Laboratory of Molecular Neuro-Oncology and Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. [5]Gene Therapy Program, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [6]Institute of Neuropathology, University of Zurich, Zurich, Switzerland. [7]Simons Foundation, New York, NY, USA. [8]Department of Computer Science, Princeton University, Princeton, NJ, USA. [9]These authors contributed equally: Jian Zhou, Christopher Y. Park, Chandra L. Theesfeld. *e-mail: darnelr@rockefeller.edu; ogt@cs.princeton.edu

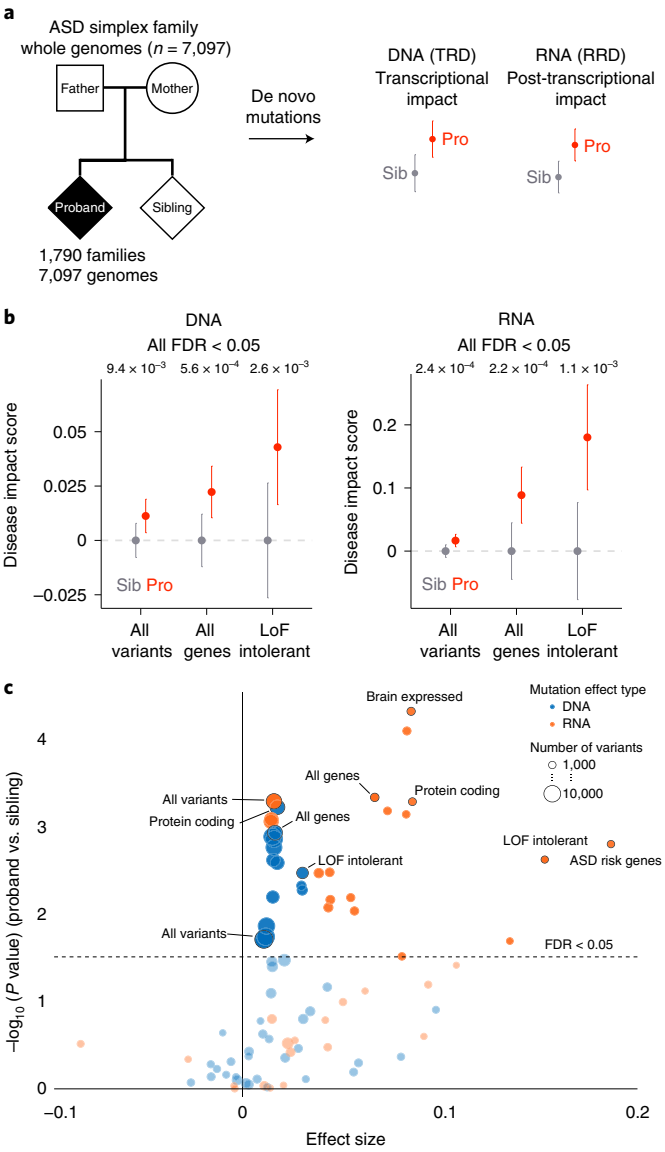

Fig. 1 | The increased effect burden of noncoding regulatory mutations in ASD. a, The overall study design for deciphering the contribution to ASD of de novo noncoding mutations genome wide. Whole genomes of 1,790 ASD simplex families were sequenced to identify de novo mutations in the ASD probands and unaffected siblings. De novo SNV mutations were analyzed for their predicted transcriptional (chromatin and transcription factors) and post-transcriptional (RBPs) regulatory effects for comparison between probands and siblings. **b**, ASD probands possess mutations with significantly higher predicted DISs as compared to their unaffected siblings. In probands, we observed a significant burden of mutations altering both transcriptional (DNA, all variants; $n=127{,}140$) and post-transcriptional (RNA, all transcribed variants; $n=77{,}149$) regulation. This proband excess was stronger when analysis was restricted to mutations near all genes for DNA ($n=69{,}328$) and near alternatively spliced exons for RNA ($n=4{,}871$), and was even stronger near ExAC LoF-intolerant genes (DNA, $n=14{,}873$; RNA, $n=1{,}355$). For analyses that included gene sets, variants were associated with the closest gene within 100 kb of the representative TSS for analysis of TRD. For analysis of RRD, variants located in introns within 400 bp of flanking exons in regions known to regulate alternative splicing were used. A Wilcoxon rank-sum test (one sided) was used for computing the significance levels. All predicted DISs were normalized by subtracting the average predicted DIS of mutations in siblings for each comparison (data are shown as mean DIS and the error bars indicate the 95% confidence interval). All results are significant after multiple-hypothesis correction (FDR < 0.05) and robust to inclusion or exclusion of mutations in protein-coding regions (Supplementary Fig. 6). **c**, Analysis with genomic variant set analysis of mutational burden for transcriptional and post-transcriptional disruptions. For each gene set and distance cutoff, the effect size (defined as the difference between the average DIS in probands and siblings) is shown on the *x* axis. A Wilcoxon rank-sum test (one sided) was used for computing the significance levels. For each category, significance levels before and after correction are listed in Supplementary Table 2. The categories shown in **b** are included in the annotation. All gene lists were obtained from Werling et al.[17]. Distance cutoffs for DNA were 10 kb, 50 kb, 100 kb, 500 kb and ∞ to TSSs; distance cutoffs for RNA were 200 bp, 400 bp and ∞ to all exons or to all alternatively spliced exons. DNA results are shown in blue and RNA results are shown in orange; dot size corresponds to sample size (number of variants in a category); total sample size $n=127{,}140$. Variant sets with more than 500 mutations are displayed. A full list of results is available in Supplementary Table 2. Uncorrected *P* values are shown on the *y* axis and the dashed line indicates categories below the FDR = 0.05 threshold after Benjamini–Hochberg correction. Results are robust to inclusion or exclusion of mutations in protein-coding regions (Supplementary Fig. 7).

factors, such as the rise in mutation counts with parental age. This difficulty is shared in other psychiatric diseases with complex genetic bases, such as intellectual disabilities and schizophrenia. In fact, little is known about the contribution of noncoding rare variants or de novo mutations to human diseases beyond the less common cases that exhibit Mendelian inheritance patterns.

To address this challenge, we used a systematic approach (Fig. 1a) that reliably identifies impactful noncoding mutations, which is analogous to using the genetic codon code to distinguish nonsynonymous mutations from synonymous mutations in protein-coding genes. This enables comparison of the mutational burden of probands and their siblings not simply in terms of the number of mutations but in terms of the functional impact of mutations. Specifically, we used biochemical data demarcating interactions between DNA- and RNA-binding proteins and their targets to train and deploy a deep convolutional-neural-network-based framework that predicts the functional and pathogenic impact of de novo mutations in the SSC using models trained for DNA and RNA. Our framework estimates, with single-nucleotide resolution, the quantitative impact of each variant on 2,002 specific transcriptional and 232 specific post-transcriptional regulatory features, including histone marks, transcription factors and RBP profiles.

Using this approach, we discovered a significantly (multiple-hypothesis-corrected) increased burden of mutations that disrupt transcriptional regulation (transcriptional-regulation-disrupting (TRD) mutations) and separately an increased burden of mutations that disrupt RBP regulation (RBP-regulation-disrupting (RRD) mutations) in ASD probands. This provides evidence of a causal role for de novo noncoding regulatory mutations in autism. Notably, the difference in functional impact between proband and sibling mutations is significant when considering de novo mutations genome wide, with increased effect sizes observed around LoF-intolerant genes (ExAC[19]). We also identify specific pathways and tissues affected by these mutations, experimentally verify the differential regulatory effect of prioritized variants and explore a link between the noncoding mutations and intelligence quotient (IQ) in ASD. We provide an interactive interface for the biomedical research community to explore the predicted impact of de novo mutations at https://hb.flatironinstitute.org/ASDbrowser/.

Results

Contribution of mutations affecting transcriptional and post-transcriptional regulation to ASD. Analysis of the contribution

of noncoding mutations to ASD is challenging due to the difficulty of assessing which noncoding mutations are functional and, of these, which contribute to the disease phenotype. To predict the regulatory impact of noncoding mutations, we constructed a deep convolutional-network-based framework to directly model the functional impact of each mutation and provide a biochemical interpretation, including disruption caused to transcription factors binding and the establishment of chromatin marks at the DNA level and to RBP binding at the RNA level (Supplementary Figs. 1 and 2). At the DNA level, the framework includes cell-type-specific models of transcriptional regulatory effects from over 2,000 genome-wide profiles of histone marks, transcription factor binding and chromatin accessibility (from the ENCODE and Roadmap Epigenomics projects[20,21]). This extends the deep-learning-based method that we described previously[10] with a redesigned architecture, leading to significantly improved performance ($P = 6.7 \times 10^{-123}$, Wilcoxon rank-sum test; Supplementary Fig. 2). At the RNA level, our deep-learning-based method was trained on the precise biochemical profiles of over 230 RBP–RNA interactions (derived from cross-linking immunoprecipitation (CLIP) data); such data can identify a wide range of post-transcriptional regulatory binding sites, including those involved in RNA splicing, localization and stability[22]. At both the transcriptional and post-transcriptional level, our models are accurate and robust in whole-chromosome-holdout evaluations (Supplementary Fig. 1b). Our models utilize a large sequence context to provide accurate single-nucleotide-resolution predictions, while also capturing dependencies and interactions between various biochemical factors (for example, histone marks or RBPs). This approach is data driven and does not rely on known sequence information, such as transcription factor-binding motifs, and it predicts the impact of any mutation regardless of whether it has been previously observed, which is essential for the analysis of de novo mutations in ASD. Finally, to link the biochemical disruption caused by a variant with phenotypic impact, we trained a regularized linear model using a set of curated regulatory noncoding mutations identified in human disease[6] (from the Human Gene Mutation Database (HGMD)) and rare variants from healthy individuals in the 1000 Genomes populations[23]. The linear model generates a predicted disease impact score (DIS) for each autism mutation independently, based on its predicted transcriptional and post-transcriptional regulatory effects.

With these approaches, we systematically assessed the functional impact of de novo mutations on the binding of regulatory factors and chromatin properties, using data derived from 7,097 whole genomes from the SSC cohort (total of 127,140 non-repeat-region single nucleotide variants (SNVs); Supplementary Table 1). When considering all de novo mutations, we observed a significantly higher functional impact in probands as compared to unaffected siblings, independently at the transcriptional level ($P = 9.4 \times 10^{-3}$, one-sided Wilcoxon rank-sum test for all; false-discovery rate (FDR) = 0.033, corrected for all mutation sets tested) and post-transcriptional level ($P = 2.4 \times 10^{-4}$, FDR = 0.0049) (Fig. 1b, all variants). This analysis is sensitive enough to discover the contribution of noncoding mutations even if only a very small fraction of the noncoding mutations are impactful (see power analysis in Supplementary Fig. 3). Furthermore, our finding is robust and significant at the level of biochemical disruptions predicted by the DNA and RNA deep-learning-based models as well as with alternative DIS training sets (Supplementary Figs. 4 and 5) or with inclusion or exclusion of protein-coding regions (Supplementary Figs. 6 and 7).

Werling et al.[17] raised the challenge of detecting significant proband-specific signal even with highly specific subsets of genes or genomic regions, and in relation to this, emphasized the need to properly correct for multiple hypotheses; this challenge was not resolved by a larger ASD cohort in a follow-up study[18]. Notably, our result does not rely on any selection of variant subsets (for example, those near predicted ASD-associated genes), is significant even after multiple-hypothesis correction and, unlike mutation counts, the predicted mutation effects are not correlated with parental age (Supplementary Fig. 8), a confounding factor of analyses based on mutation counts.

To gain further insight into the noncoding regulatory landscape in ASD, we conducted a comprehensive analysis with full multiple-hypothesis correction for all 140 combinations of the 14 gene sets previously used in Werling et al.[17], examined across ten genomic regions (for example, transcription start site (TSS)-proximal regions and exon-proximal regions). When analysis was restricted to genomic regions of higher regulatory potential (that is, near TSSs or alternatively spliced exons), we observed an increased effect size for dysregulation (Fig. 1b,c; all genes; TRD: $P = 5.6 \times 10^{-4}$, FDR = 0.0056; RRD $P = 2.2 \times 10^{-4}$, FDR = 0.0048). Among gene sets, we observed an increased proband burden of high-effect mutations close to LoF-intolerant genes (probablity of being LoF intolerant (pLI) > 0.9 from ExAC; 3,230 genes; TRD: $P = 2.6 \times 10^{-3}$, FDR = 0.013; RRD: $P = 1.1 \times 10^{-3}$, FDR = 0.0078) (Fig. 1b,c and Supplementary Fig. 9). This finding suggests that, in ASD, LoF-intolerant genes are highly vulnerable to noncoding disruptive mutations. This is consistent with the enrichment of coding LoF mutations among LoF-intolerant genes in the SSC cohort[24], indicating ASD signal convergence of noncoding and coding de novo mutations. Furthermore, we also found convergent signal at both the transcriptional and post-transcriptional level, thus providing further evidence for a causal role of noncoding effects in ASD (a full list of *P* values and FDRs is available in Supplementary Table 2). We observed these signals consistently across the SSC cohort subsets that were sequenced in different phases (Supplementary Fig. 10).

Tissue specificity and functional landscape of de novo ASD-associated noncoding mutations. Although one of the hallmarks of autism is altered brain development, a comprehensive tissue association has not been established for de novo noncoding variants. To explore the proband-specific signal in different tissues, we systematically tested the variant effects for genes with tissue-specific expression derived for all 53 tissues and cell types in the Genotype–Tissue Expression (GTEx) project[25]. We observed a consistent significant proband-specific mutation effect associated with brain tissues, with brain regions constituting the top 11 most highly ranked tissues (ranked by the difference in the effect of noncoding mutations in proband versus sibling) (Fig. 2a; all with FDR < 0.05). This provides strong evidence that high-impact variants from the noncoding genome of ASD probands likely disrupt brain-specific gene regulation, which is consistent with previous findings for mutations in protein-coding regions[26].

We next investigated the underlying processes and pathways impacted by de novo noncoding mutations in ASD. Such analysis is challenging because, in addition to the variability in the functional impact of mutations, ASD probands appear highly heterogeneous in underlying causal genetic perturbations[27] and single mutations could cause a widespread effect on downstream genes. Thus, to detect genes and pathways relevant to the pathogenicity of TRD and RRD mutations in ASD, we developed a network-based statistical approach, which we term network-neighborhood differential enrichment analysis (NDEA; Supplementary Fig. 11). We used a brain-specific functional network that probabilistically integrates a large compendium of public omics data (for example, expression, protein–protein interaction (PPI) and motifs) to represent how likely it is that two genes act together in a biological process[28]. When applied to ASD de novo mutations, the NDEA approach identifies genes whose functional network neighborhood is significantly enriched for genes with stronger predicted disease impact in proband mutations as compared to sibling mutations (Supplementary Table 3).

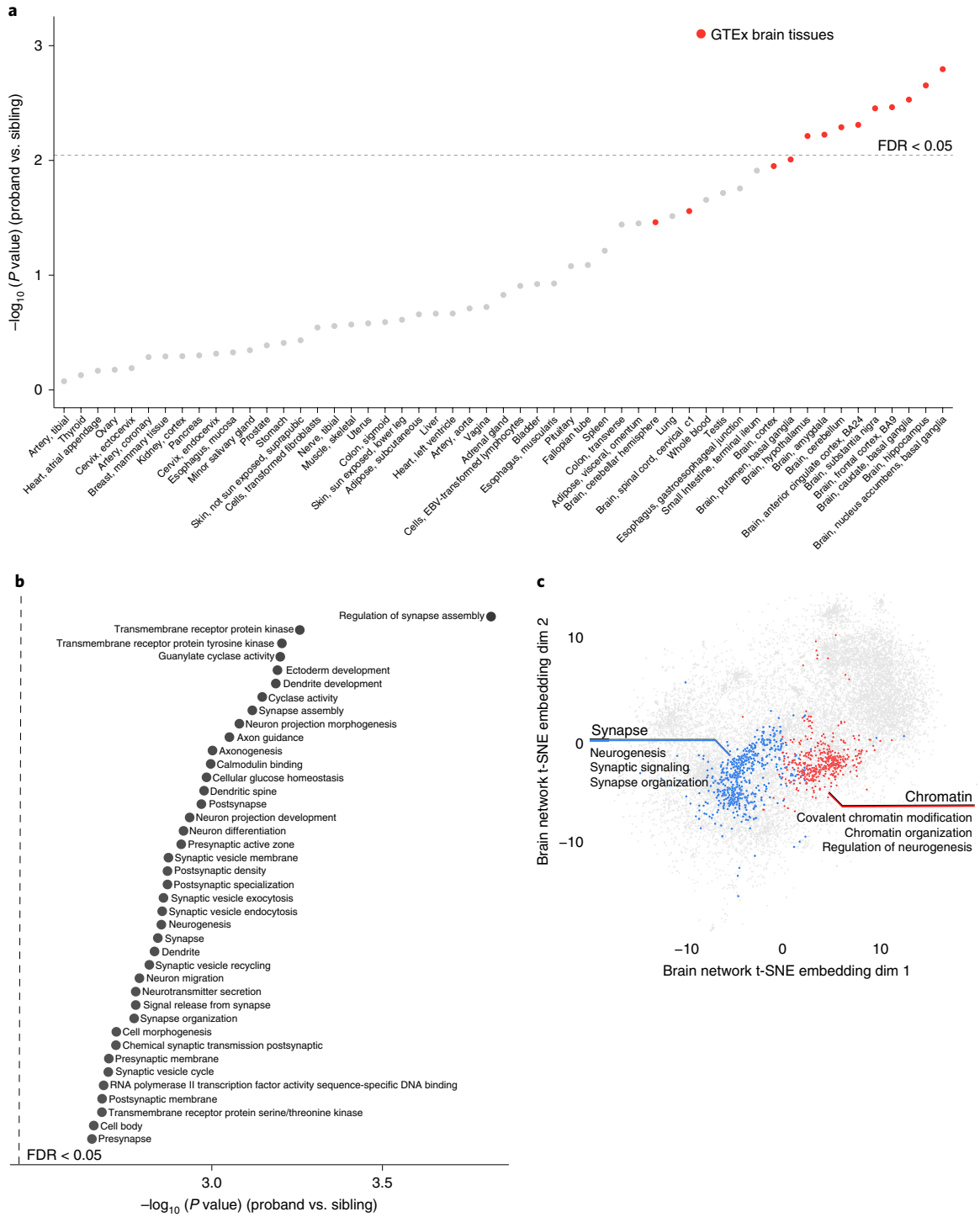

Fig. 2 | Analysis of the effects of noncoding mutations converges on brain-specific signals and neurodevelopmental processes. a, Brain-tissue-specific genes show the strongest elevated proband-specific noncoding-mutation effect burden. All 53 GTEx tissues are ranked by significance of increased mutation burden in genes with tissue-specific expression in probands as compared to unaffected siblings (Methods). Uncorrected *P* values are shown on the *y* axis and the dashed line indicates tissues below the FDR = 0.05 threshold after Benjamini–Hochberg correction. DISs for all mutations within 100 kb of representative TSSs (DNA) and intronic mutations within 400 bp of exon boundaries (RNA) were used for the analysis ($n = 71{,}554$). **b**, Processes related to neuronal function and development show a significant excess of proband mutation DISs as demonstrated by NDEA (full list in Supplementary Table 4 also see the Methods). Analysis was conducted on the same mutation set as in **a**. The top processes (*y* axis) and the *P* values of proband excess (*x* axis) are shown. Uncorrected *P* values are shown on the *x* axis and all gene sets shown have FDR < 0.05. **c**, Genes with significant network neighborhood excess of high-impact proband mutations form two functionally coherent clusters (see annotations for representative enriched gene sets in each cluster; a full list is available in Supplementary Table 5). Analysis was conducted on the same mutation set as in **a**. The brain functional network was visualized by computing two-dimensional embeddings with t-SNE (Methods). Genes, but not network edges, are shown for clarity of visualization. Clustering was performed with Louvain community clustering. All genes in the two clusters shown have FDR < 0.1.

Globally, NDEA enrichment analysis pointed to a proband-specific role for noncoding-mutation effects in neuronal development, including in synaptic transmission and chromatin regulation (Fig. 2b and Supplementary Table 4), consistent with processes that have been previously associated with ASD, based on protein-coding variants[2,26]. Genes with significant NDEA enrichment were specifically involved in neurogenesis and grouped into two functionally coherent clusters using the Louvain community-detection algorithm (Fig. 2c and Supplementary Table 5). The synaptic cluster is enriched in ion channels and receptors involved in neurogenesis ($P=5.6\times10^{-38}$), synaptic signaling ($P=4.8\times10^{-35}$) and synapse organization ($P=1.5\times10^{-18}$), including previously known ASD-associated genes such as those involved in synapse organization (*SHANK2*, *NLGN2* and *NRXN2*), synaptic signaling (*NTRK2* and *NTRK3*), ion channels (*CACNA1A*, *CACNA1C*, *CACNA1E*, *CACNA1G* and *KCNQ2*) and neurotransmission (*SYNGAP1*, *GABRB3*, *GRIA1* and *GRIN2A*)[29]. The synapse cluster is also significantly enriched for plasma membrane proteins ($P=3.9\times10^{-24}$). In contrast, the chromatin cluster, representing processes related to chromatin regulation, displayed an over-representation of nucleoplasm proteins ($P=2.1\times10^{-9}$), with diverse functional roles including covalent chromatin modification ($P=2.5\times10^{-9}$), chromatin organization ($P=5.2\times10^{-8}$) and regulation of neurogenesis ($P=6.4\times10^{-5}$). The chromatin cluster also includes many known ASD-associated genes such as the chromatin remodeler *CHD8*, the chromatin modifiers *KMT2A* and *KDM6B* and the Parkinson's disease gene *PINK1* (ref. [30]), which is also associated with ASD[29] (Supplementary Table 3). Overall, our results demonstrate pathway-level TRD and RRD mutation burden and identify distinct network-level hot spots for high-impact de novo mutations.

Next, we examined the genetic landscape of ASD-associated de novo noncoding and coding mutations. Specifically, in addition to the network analysis of noncoding mutations at the transcriptional and post-transcriptional level, we also applied network analysis to de novo coding mutations[2]. We compared the gene-specific NDEA statistic of the proband-specific effect burden for noncoding mutations with that of coding mutations, finding a significant positive correlation for both TRD and RRD ($P=0.004$ and Pearson's $r=0.39$ for TRD; $P=0.042$ and Pearson's $r=0.30$ for RRD; two-sided permutation test). Moreover, network analysis showed that TRD and RRD are themselves significantly correlated ($P=0.034$, and Pearson's $r=0.36$; two-sided permutation test). This demonstrates that coding and noncoding mutations affect overlapping processes and pathways, which indicates a convergent genetic landscape and highlights the potential for the discovery of ASD-associated genes by combining coding and noncoding mutations.

Experimental study of the effects of ASD-associated noncoding mutations on gene regulation. Our analysis identified new candidate noncoding disease-associated mutations that potentially affect ASD through regulation of gene expression. To add further evidence to a set of high-confidence causal mutations, we experimentally studied the allele-specific effects of predicted high-impact mutations in cell-based assays. For TRD mutations, 59 genomic regions exhibited strong transcriptional activity, with 96% of proband variants (57 variants) showing robust differential activity (Fig. 3 and Methods), demonstrating that our prioritized de novo TRD mutations do indeed lie in regions with transcriptional regulatory potential and that the predicted effects translate to measurable allele-specific effects on expression. Among the genes with mutations that exhibited strong differential activity were *NEUROG1*, which encodes an important regulator of initiation of neuronal differentiation, and *DLGAP2*, which encodes a guanylate kinase that is localized to the postsynaptic density in neurons. In the NDEA analysis, *NEUROG1* had significant network-neighborhood excess in probands ($P=8.5\times10^{-4}$). Mutations near *HES1* and *FEZF1* also had significant differential effects on activator activities. Neurogenin, HES and FEZF family transcription factors act together during development, both receiving and sending inputs to Wnt and Notch signaling in the developing central nervous system and, interestingly, in the gut to control stem cell fate decisions[26,31–34], and Wnt and Notch pathways have been previously associated with autism[27,35]. *SDC2* encodes a synaptic syndecan protein involved in the formation of dendritic spines and synaptic maturation, and a structural variant near the 3′ end of the gene was reported in an autistic individual (reviewed in ref. [36]). Thus, our method identified alleles of high predicted impact that do indeed result in changes in transcriptional regulatory activity in cells. As many autism genes are under strong evolutionary selection, only effects exerted through (more subtle) changes in gene expression may be observable because complete LoF mutations may be lethal. This implies that further study of the prioritized noncoding regulatory mutations should yield insights into the range of dysregulation associated with autism.

In addition, as a case study for prioritized RRD mutations, we experimentally validated the effect of an ASD proband de novo noncoding mutation lying outside of a canonical splice site that we predicted to disrupt splicing of *SMEK1* (ExAC pLI = 1.0; Supplementary Fig. 12). Smek1 has previously been shown to regulate cortical neurogenesis through the Wnt signaling pathway[37]. For this mutation, we observed a reduction of more than 40% in the inclusion of the exon for the ASD proband allele as compared to the sibling allele in a minigene assay (Methods), in agreement with the high predicted RRD impact. This demonstrates the highly disruptive biochemical impact that a de novo mutation outside a canonical splice site can have on RNA splicing.

The individual-level clinical relevance of the noncoding de novo mutations. The majority of ASD probands in the SSC do not have a de novo LoF coding mutation[1,2], and noncoding mutations outnumber LoF coding mutations by over 500-fold[18]. While the individual effect of noncoding mutations may vary, as a group, noncoding mutations could have significant clinical impact. Indeed, we observed a significant increase in ASD risk for individuals with a higher burden of impactful de novo mutations (Supplementary Fig. 9; mean DIS per individual, Wilcoxon rank-sum test one-sided $P=1.4\times10^{-3}$), with 25% of the SSC ASD probands incurring an aggregate noncoding ASD risk of 1.2 (odds ratio).

Furthermore, the overall contribution of de novo noncoding mutations (which explain 4.3% of the SSC ASD cases) was comparable to that of LoF coding mutations (5.4%) and to that of missense mutations (3.1%) (Supplementary Fig. 13). This analysis leverages the power of the quad simplex design of the SSC cohort, enabling the estimation of the causal contribution of each mutation category by correcting for the background occurrence rate among unaffected siblings (Methods). Thus, our results demonstrate that noncoding de novo mutations have clinical relevance, although not all ASD probands will have impactful noncoding mutations (even in aggregate), and future work will be required to characterize their clinical impact and relationship to phenotypes.

One interesting direction is linking the effects of noncoding mutations to specific phenotypes, such as IQ heterogeneity among ASD probands. Intellectual disability is estimated to impact 40–60% of children with ASD[38] and individuals with ASD can over-inherit common variants associated with educational attainment[39]. For de novo noncoding mutations analyzed in this study, we observed a significant association between noncoding mutations and IQ in individuals with ASD. Specifically, individuals with ASD with lower IQ had a higher burden of RRD mutations in intronic regions flanking alternatively spliced exons of ExAC LoF-intolerant genes. This provides genetic evidence that aberrant splicing can contribute to the phenotypic heterogeneity observed among ASD probands (Supplementary Fig. 14; $P=1.5\times10^{-3}$) and should be taken into account when projecting clinical outcomes.

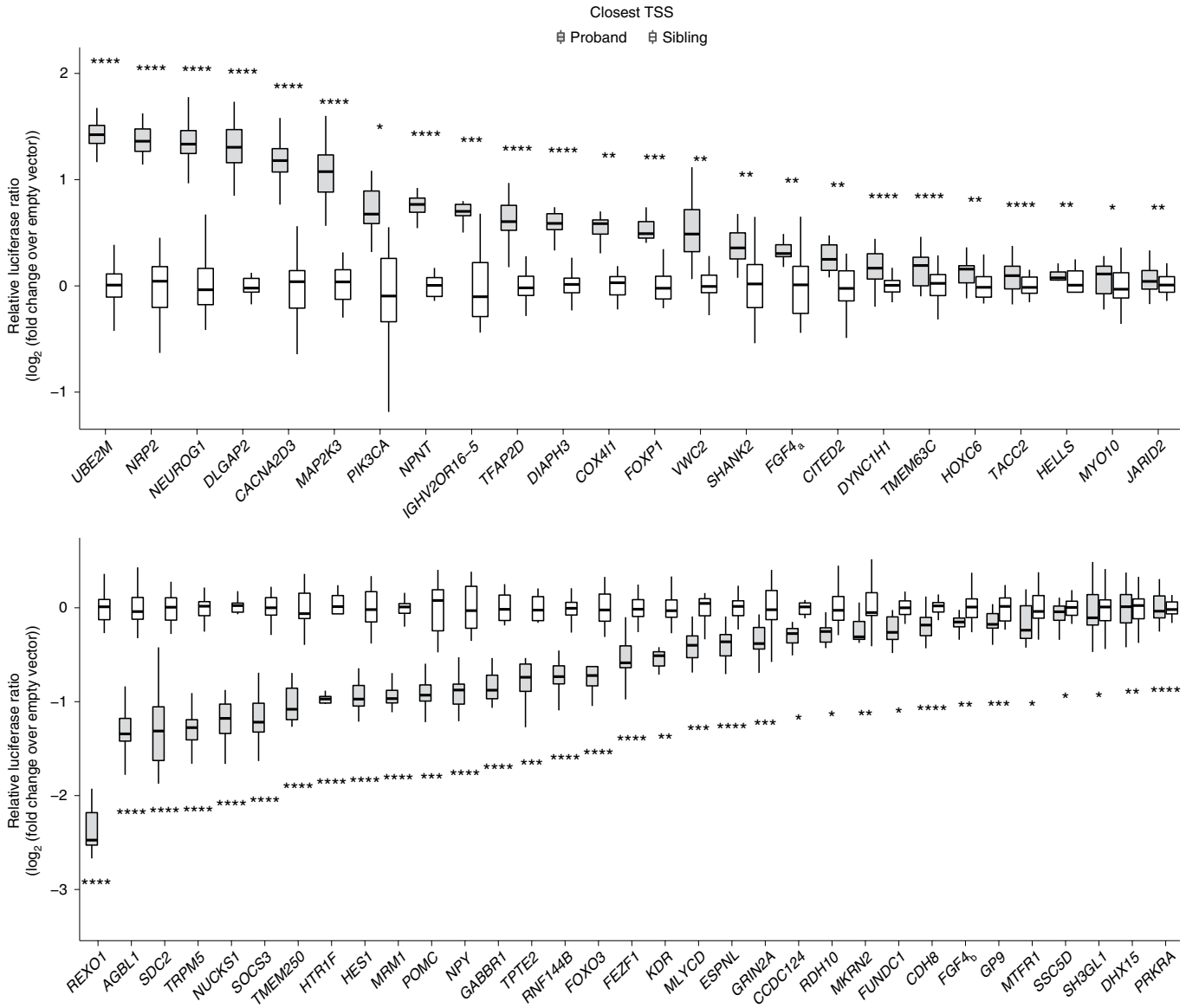

Fig. 3 | Allele-specific transcriptional activity of ASD noncoding mutations. Differential expression by proband and sibling alleles in a dual-luciferase assay demonstrates that 57 predicted high-impact TRD mutations associated with ASD fall in active regulatory elements and the mutations confer substantial changes to the regulatory potential of the sequence. Cells were transfected with a transfection control and a pGL4.23-based expression plasmid containing 230 nucleotides of the genomic region, and luminescence was assayed 42 h later (Methods). The *y* axis shows the magnitude of transcriptional activation normalized to the activity for the sibling allele. Significance levels were computed on the basis of a *t* test and Fisher's combined probability test (two sided; $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, $^{****}P<0.0001$; Methods). Sample sizes for all tests are listed in Supplementary Table 6. Central values of the box plot represent the median, the box extends from the twenty-fifth to the seventy-fifth percentile, and whiskers extend to the maximum and minimum values no further than 1.5 times the interquartile range from the hinge.

Discussion

Even with the great strides made in understanding the causes of ASD by sequencing and phenotyping of multiple cohorts in recent years, much of the genetic basis underlying autism remains undiscovered. While a number of coding variants have been associated with ASD, no systematic evidence of de novo noncoding effects has been observed. Here we present a new deep-learning-based approach for quantitatively assessing the impact of noncoding mutations on human disease. Our approach addresses the statistical challenge of detecting the contribution of noncoding mutations by predicting their specific effects on transcriptional and post-transcriptional regulation. This approach is general and can be applied to study the contributions of noncoding mutations to any complex disease or phenotype.

Here we apply our strategy to ASD using the 1,790 whole-genome-sequenced families from the SSC and, to our knowledge, demonstrate for the first time significant proband-specific signal in regulatory de novo noncoding space. Importantly, we not only detect this signal at the transcriptional level but also independently find significant proband-specific RRD burden. Previously, there has been limited evidence for disease contribution of mutations disrupting post-transcriptional mechanisms outside of canonical splice sites. We demonstrate significant ASD disease association at the level of de novo mutations for variants impacting a large collec-

tion of RBPs regulating post-transcriptional regulation. Overall, our results suggest that both transcriptional and post-transcriptional mechanisms play a major role in the etiology of ASD and possibly other complex diseases.

Notably, our study reveals important biological convergences among the genetic dysregulations associated with ASD. Our analyses of the disease impact of mutations with effects on DNA and RNA point to similar sets of impacted genes and pathways, indicating that the effects of regulatory mutations are convergent. Furthermore, high-impact noncoding regions that we find in ASD probands affect the same genes previously found to be impacted by LoF coding mutations in ASD. This convergence provides support for a causal contribution of noncoding regulatory mutations to ASD etiology.

Our analyses also demonstrate the potential for predicting disease phenotypes from genetic information, including de novo noncoding mutations. We provide a resource for further research into understanding the mechanism of noncoding effects on ASD, including computationally prioritized TRD and RRD mutations with strong predicted regulatory effects, as well as ASD proband mutations that potentially contribute to disease with experimentally confirmed effects (Supplementary Tables 1 and 6; https://hb.flatironinstitute.org/ASDbrowser/). However, there remains much room for further progress in this important area. We expect that continuing development of methods for predicting the effects of noncoding mutations will further improve the power of WGS studies for discovering the biological mechanisms of the contributions of noncoding mutations to autism and other complex human diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0420-0.

Received: 3 August 2018; Accepted: 12 April 2019;
Published online: 27 May 2019

References

1. Sanders, S. J. et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
2. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
3. Yuen, R. K. C. et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat. Neurosci.* **20**, 602–611 (2017).
4. Bernstein, B. E. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
5. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
6. Stenson, P. D. et al. The human gene mutation database: 2008 update. *Genome Med.* **1**, 13 (2009).
7. Feigin, M. E. et al. Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma. *Nat. Genet.* **49**, 825–833 (2017).
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. Brandler, W. M. et al. Paternally inherited *cis*-regulatory structural variants are associated with autism. *Science* **360**, 327–331 (2018).
10. 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).
11. Turner, T. N. et al. Genomic patterns of de novo mutation in simplex autism. *Cell* **171**, 710–722 (2017).
12. Yuen, R. K. C. et al. Genome-wide characteristics of de novo mutations in autism. *NPJ Genom. Med.* **1**, 16027 (2016).
13. Yuen, R. K. C. et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.* **21**, 185–191 (2015).
14. Michaelson, J. J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**, 1431–1442 (2012).
15. Jiang, Y. et al. Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am. J. Hum. Genet.* **93**, 249–263 (2013).
16. Kong, A. et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
17. Werling, D. M. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* **50**, 727–736 (2018).
18. An, J. Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).
19. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
20. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
21. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
22. Ule, J., Hwang, H.-W. & Darnell, R. B. The future of cross-linking and immunoprecipitation (CLIP). *Cold Spring Harb. Perspect. Biol.* **10**, a032243 (2018).
23. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
24. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
25. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
26. Packer, A. Neocortical neurogenesis and the etiology of autism spectrum disorder. *Neurosci. Biobehav. Rev.* **64**, 185–195 (2016).
27. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
28. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
29. Iossifov, I. et al. Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl Acad. Sci. USA* **112**, E5600–E5607 (2015).
30. Valente, E. M. Hereditary early-onset Parkinson's disease caused by mutations in *PINK1*. *Science* **304**, 1158–1160 (2004).
31. Kageyama, R. & Ohtsuka, T. The Notch–Hes pathway in mammalian neural development. *Cell Res.* **9**, 179–188 (1999).
32. Bertrand, N., Castro, D. S. & Guillemot, F. Proneural genes and the specification of neural cell types. *Nat. Rev. Neurosci.* **3**, 517–530 (2002).
33. Crosnier, C., Stamataki, D. & Lewis, J. Organizing cell renewal in the intestine: stem cells, signals and combinatorial control. *Nat. Rev. Genet.* **7**, 349–359 (2006).
34. Eckler, M. J. & Chen, B. Fez family transcription factors: controlling neurogenesis and cell fate in the developing mammalian nervous system. *BioEssays* **36**, 788–797 (2014).
35. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Res.* **25**, 142–154 (2015).
36. Saied-Santiago, K. & Blow, H. E. Diverse roles for glycosaminoglycans in neural patterning. *Dev. Dyn.* **247**, 54–74 (2017).
37. Chang, W.-H. et al. Smek1/2 is a nuclear chaperone and cofactor for cleaved Wnt receptor Ryk, regulating cortical neurogenesis. *Proc. Natl Acad. Sci. USA* **114**, E10717–E10725 (2017).
38. Walsh, C. A., Morrow, E. M. & Rubenstein, J. L. R. Autism and brain development. *Cell* **135**, 396–400 (2008).
39. Weiner, D., Wigdor, E., Ripke, S. & Robinson, E. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat. Genet.* **49**, 978–985 (2017).

Acknowledgements

We are grateful to the families participating in the SFARI SSC. This work is supported by NIH grants R01HG005998, U54HL117798 and R01GM071966, HHS grant HHSN272201000054C and Simons Foundation grant 395506 to O.G.T.; NIH grants 1UM1HG008901, NS034389, NS081706 and NS097404 and Simons Foundation grant SFARI 240432 to R.B.D.; and STARR Cancer Consortium Award I10-0056 to C.Y.P. and R.B.D. O.G.T. is a senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research (CIFAR). R.B.D. is an Investigator of the Howard Hughes Medical Institute. The authors acknowledge all members of the Troyanskaya and Darnell laboratory for helpful discussions. We also thank the SFARI, Simons Foundation and Flatiron Institute, in particular N. Volfovsky and M. Benedetti. We are pleased to acknowledge that a substantial portion of the work in this paper was performed at the TIGRESS high-performance computer center at Princeton University, which is jointly supported by the Princeton Institute for Computational Science and Engineering and the Princeton University Office of Information Technology's Research Computing department. O.G.T. is a CIFAR fellow.

Author contributions

J.Z., C.Y.P., C.L.T., R.B.D. and O.G.T. conceived and designed the study. J.Z. and C.Y.P. developed the computational methods and performed the analyses. J.Z. developed the DNA model and C.Y.P. developed the RNA model. C.L.T. designed and performed luciferase assay experiments. Y.Y., C.S., J.J.F. and Y.T. designed and performed the minigene splicing assay and RBP experiments. A.K.W., J.F. and K.Y. developed the web interface. A.P. contributed ideas and insights. J.Z., C.Y.P., C.L.T., R.B.D. and O.G.T. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0420-0.

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

Correspondence and requests for materials should be addressed to R.B.D. or O.G.T.

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

Methods

De novo mutation calling and filtering. The SSC WGS data were made available via the Simons Foundation Autism Research Initiative (SFARI) and were processed to generate variant calls via the standard GATK pipeline. The SSC WGS data can be requested through SFARI Base (https://www.sfari.org/resource/sfari-base/), with the condition that the use of the data is limited to projects related to advancing the field of autism and related neurodevelopmental disorder research (questions on SSC consents should be directed to collections@sfari.org). To call de novo single-nucleotide substitutions, inherited mutations were removed and candidate de novo mutations were selected from the GATK variant calls where the alleles were not present in parents and the parents were homozygous for the same allele. The DNMFilter[40] classifier was then used to score each candidate de novo mutation; a threshold of probability > 0.75 was applied for SSC phases 1–2 and a threshold of probability > 0.5 was applied for phase 3 to obtain a comparable number of high-confidence de novo mutation calls across phases.

The DNMFilter[40] classifier was trained with an expanded training set combining the original training standards with the verified de novo mutations from the SSC pilot WGS studies for the initial 40 SSC families. For final analysis, de novo mutation calls within the low-complexity repeat regions from the UCSC browser table RepeatMasker[41] were removed. Also, de novo mutations appearing in multiple SSC families (that is, non-singleton de novo mutations) or individuals with outlier numbers of mutations (>3 s.d. above the average) were excluded from the analysis.

Overall across the genome, we detected 77.7 mutations per individual with a transition-to-transversion (Ti/Tv) ratio of 2.01 (95% confidence interval (2.00, 2.03)) (78.7 for probands with Ti/Tv = 2.02 (1.99, 2.04), 76.7 for siblings with Ti/Tv = 2.01 (1.99, 2.03)) and no significant difference in mutation substitution patterns between probands and siblings (Supplementary Fig. 15). The WGS de novo mutation calls were compared with de novo mutations calls from exome sequencing and previously validated SSC de novo mutations[15]: 87.9% of the mutation calls from exome sequencing and 90.3% of the validated mutations were rediscovered in our mutation calls.

Training models of DNA transcriptional regulatory effects and RNA post-transcriptional effects. For training the transcriptional regulatory effects model, training labels, such as histone marks, transcription factors and DNase I profiles, were processed from uniformly processed ENCODE and Roadmap Epigenomics data releases. The training procedure is as described in Zhou and Troyanskaya[21] with the following modifications. The model architecture was extended to double the number of convolution layers for increased model depth (Supplementary Note). Similarly to our previous model[21], all layers except for the last linear layer were shared across all biochemical features. Input features were expanded to include all of the released Roadmap Epigenomics histone marks and DNase I profiles, resulting in 2,002 total features (Supplementary Table 7), as compared with the original 919 features.

For training the post-transcriptional regulatory effects model, we utilized the DeepSEA network architecture and training procedure with RBP profiles as training labels (a full list of parameters used in the model is in the Supplementary Note). We uniformly processed RNA features composed of 231 CLIP binding profiles for 82 unique RBPs (ENCODE and previously published CLIP datasets) and a branchpoint mapping profile as input features (a full list of experimental features appears in Supplementary Table 8). CLIP data processing followed our previously detailed pipeline[42], and all CLIP peaks with $P < 0.1$ were used for training with an additional filter requirement of twofold enrichment over input for ENCODE eCLIP data. In contrast to DeepSEA, only transcribed genic regions were considered as training labels for the post-transcriptional regulatory effects model. Specifically, all gene regions defined by Ensembl (mouse build 80, human build 75) were split into 50-nucleotide bins in the transcribed strand sequence. For each sequence bin, RBP profiles that overlapped more than half were assigned a positive label for the corresponding RBP model. Negative labels for a given RBP model were assigned to sequence bins where non-overlapping peaks of other RBPs were observed. Our deep learning models, both transcriptional and post-transcriptional, do not use any mutation data for training; thus, the models can predict impacts for any mutation regardless of whether it has been previously observed.

Prediction of disease impact scores. We used curated disease-associated mutations in regulatory regions and rare variants from healthy individuals to train a model that prioritizes likely disease-associated mutations on the basis of the predicted transcriptional or post-transcriptional regulatory effects of these mutations. As positive examples, we used 4,401 regulatory noncoding mutations curated in the HGMD with mutation type 'regulatory', including the sub-categories disease-causing mutation (DM), disease-causing mutation? (DM?), disease-associated polymorphism with supporting functional evidence (DFP), disease-associated polymorphism (DP) and in vitro/laboratory or in vivo functional polymorphism (FP). For negative examples of background mutations, we used 999,668 rare variants that were only observed once within the healthy individuals from the 1000 Genomes project[23]. We also showed that using an alternative set of negative variants gives similar conclusions: common variants with allele frequency > 0.01 and located within 100 kb to positives (HGMD regulatory variants). (Supplementary Fig. 5). Absolute differences in predicted probability computed by the convolutional network model of transcriptional regulatory effects (described above) were used as input features for each of the 2,002 transcriptional regulatory features and for the 232 post-transcriptional regulatory features in the model of disease impact. Input features were standardized to unit variance and zero mean before being used for training. We separately trained an L2 regularized logistic regression model for the model of transcriptional effects ($\lambda = 10$) and the model of post-transcriptional effects ($\lambda = 10$, using only examples of genic region variants) with the xgboost package (https://github.com/dmlc/xgboost/). The positive and negative training samples were separately weighted according to the inverse of the number of samples to address the label imbalance. The predicted probabilities were z transformed to have mean = 0 and s.d. = 1 across all proband and sibling mutations.

Gene sets and resources. All gene sets used are from Werling et al.[17]. The 14 gene sets include GENCODE protein-coding genes, antisense genes, long intergenic noncoding RNA genes, pseudogenes, genes with pLI > 0.9 from ExAC[19], predicted ASD risk genes (FDR < 0.3) from Sanders et al.[8], target genes of the fragile X mental retardation protein[43], genes associated with developmental delay[44,45] and CHD8 target genes[46,47]. For genes with expression specific to each of the 53 GTEx tissues, we used the expression table from GTEx v.7 (gene median transcripts per million (TPM) per tissue)[25] and we selected genes for which expression in a given tissue was five times higher than the median expression across all tissues.

We determined the representative TSS for each gene on the basis of FANTOM CAGE transcription initiation counts relative to GENCODE gene models. Specifically, a CAGE peak is associated with a GENCODE gene if it is within 1,000 bp of a GENCODE v.24 annotated TSS[48,49]. Peaks within 1,000 bp of rRNA, small nuclear RNA, small nucleolar RNA or tRNA genes were removed to avoid confusion. Next, we selected the most abundant CAGE peak for each gene and took the TSS position reported for the CAGE peak as the selected representative TSS for the gene. For genes with no CAGE peaks assigned, we kept the GENCODE annotated gene start position as the representative TSS. FANTOM CAGE peak abundance data were downloaded at http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/ and the CAGE read counts were aggregated over all FANTOM 5 tissue and cell types. GENCODE v.24 annotations lifted to GRCh37 coordinates were downloaded from https://www.gencodegenes.org/human/release_24lift37.html. All chromatin profiles used from the ENCODE and Roadmap Epigenomics projects are listed in Supplementary Table 7. The HGMD mutations are from HGMD professional v.2018.1.

Human exons that are alternatively spliced were obtained from a recent study that examined publicly available human RNA-seq data to annotate an extensive catalog of alternative splicing events[50]. Internal exon regions (both 5′ splice site (SS) and 3′ SS flanking introns), the upstream exon (5′ SS flanking intron) and the downstream terminal exon (3′ SS flanking intron) were used for definition of alternative exon types: cassette, mutually exclusive and tandem cassette exons. The terminal exon region was used for intron retention and alternatively spliced 3′ or 5′ exon types. All selected exon-flanking intronic regions were collapsed into a final set of genomic intervals used to subset SNVs that were located within the alternative splicing exon regions (200 or 400 nucleotides from the exon boundary; Supplementary Fig. 16).

Network differential enrichment analysis. Networks of brain-specific functional relationships integrate a wide range of functional genomic data in a tissue-specific manner and predict the probability of functional association between any pair of genes[28]. This network was filtered to only include edges with probability > 0.01 (above the Bayesian prior) to reduce the impact of noisy low-confidence edges.

We used NDEA to test the differential (proband versus sibling) impact of mutations on each gene or gene set. Intuitively, this test generates a P value that reflects the proband-specific impact of mutations on that gene or gene set, including through its network neighborhood. This also enables statistical assessment of which gene sets (pathways) are significantly more affected by proband mutations than sibling mutations. Technically, NDEA performs a weighted two-sample (proband versus sibling mutations) test, where the weight for each observation is defined on the basis of the network connectivity scores (to the gene or gene sets) and the weighted averages of two samples are compared. Each weight is a non-negative constant number that is used to specify the relative contribution of an observation to the test statistic. When all weights are the same, it reduces to regular two-sample t tests; when the weights are different, the standard t statistic is adjusted to use appropriate variance according to weighting. Unlike some other forms of weighted t test, the weights are not random variables and do not represent sample sizes. The assumptions of the NDEA test are analogous to those of the standard two-sample t test, including that samples in each set are independent and identically distributed random variables and that the weighted sample means are normally distributed.

For each gene i, the NDEA t statistic is computed by

$$t_i = (\mu_{P_i} - \mu_{S_i}) / \sqrt{\frac{V_{P_i}}{N_{P_i}} + \frac{V_{S_i}}{N_{S_i}}}$$

$$\mu_{P_i} = \frac{\sum_{m \in P} W_{ij(m)} d_m}{\sum_{m \in P} W_{ij(m)}}$$

$$\mu_{S_i} = \frac{\sum_{m \in S} W_{ij(m)} d_m}{\sum_{m \in S} W_{ij(m)}}$$

$$V_{P_i} = \frac{\sum_{m \in P} W_{ij(m)} (d_m - \mu_{P_i})^2}{\sum_{m \in P} W_{ij(m)} - \frac{\sum_{m \in P} W_{ij(m)}^2}{\sum_{m \in P} W_{ij(m)}}}$$

$$V_{S_i} = \frac{\sum_{m \in S} W_{ij(m)} (d_m - \mu_{S_i})^2}{\sum_{m \in S} W_{ij(m)} - \frac{\sum_{m \in S} W_{ij(m)}^2}{\sum_{m \in S} W_{ij(m)}}}$$

$$N_{P_i} = \frac{\left(\sum_{m \in P} W_{ij(m)}\right)^2}{\sum_{m \in P} W_{ij(m)}^2}, N_{S_i} = \frac{\left(\sum_{m \in S} W_{ij(m)}\right)^2}{\sum_{m \in S} W_{ij(m)}^2}$$

in which μ_{P_i} and μ_{S_i} are weighted averages of DIS d_m of all proband mutations P or all sibling mutations S. V_{P_i} and V_{S_i} are the unbiased estimates of population variance of μ_{P_i} and μ_{S_i}. N_{P_i} and N_{S_i} are the effective sample sizes of proband and sibling mutations after network-based weighting for gene i. $W_{ij(m)}$ is the network edge score (interpreted as the functional relationship probability) between gene i and gene $j(m)$ divided by the number of proband (if m is a proband mutation) or sibling (if m is a sibling mutation) mutations that gene $j(m)$ is associated with, where $j(m)$ indicates the implicated gene of the mutation m. P and S are the set of all proband mutations and the set of all sibling mutations included in the analysis.

Under the null hypothesis of the two groups having no difference, the above t statistic approximately follows a t distribution with the following degree of freedom.

$$df = \frac{\left(\frac{V_{P_i}}{N_{P_i}} + \frac{V_{S_i}}{N_{S_i}}\right)^2}{\frac{V_{P_i}^2}{N_{P_i}^2 (N_{P_i} - 1)} + \frac{V_{S_i}^2}{N_{S_i}^2 (N_{S_i} - 1)}}$$

For testing the significance of differences between proband and sibling mutations, mutations within 100 kb of the representative TSS of all genes and all intronic mutations within 400 bp of an exon boundary were included in this analysis. RNA DISs were used as the mutation score for intronic mutations within 400 bp of an exon boundary and DNA DISs were used for other mutations.

For NDEA, at the gene set level, we consider the gene set as a meta-node that contains all genes that are annotated to the gene set (for example, a Gene Ontology (GO) term). Then, for any given gene, the average of network edge scores for all genes in the meta-node is used as the weights. GO term annotations were pooled from human (EBI, 9 May 2017), mouse (MGI, 26 May 2017) and rat (RGD, 8 April 2017). Query GO terms were obtained from the merged set of curated GO Consortium[51] slims from Generic, Synapse and ChEMBL and supplemented by PANTHER[52] GO-slim and terms from NIGO[53].

For network-based analysis of correlation between mutations in protein-coding regions and noncoding TRD and RRD mutations, we first compute the NDEA t statistic for every gene for all mutations in protein-coding regions from the SSC exome sequencing study[2,8], all SSC WGS noncoding mutations within 100 kb of a gene and all SSC WGS genic noncoding mutations within 400 bp of an exon, respectively. We then compute Pearson correlation across all resulting gene-specific t statistics for all three pairs of mutation types. For testing statistical significance of the correlation, we permuted proband and sibling labels for all mutations to compute the null distributions of correlations for each pair of mutation types. One thousand permutations were performed.

Network visualization and clustering. For network visualization, we computed a two-dimensional embedding with t-distributed stochastic neighbor embedding (t-SNE)[54] by directly taking a distance matrix of all pairs of genes as the input. The distance matrix was computed as −log(probability) from the edge probability score matrix in the brain-specific functional relationship network. The Barnes–Hut t-SNE algorithm implemented in the Rtsne package was used for the computation. Louvain community clustering was performed on the subnetwork containing all protein-coding genes with the top 10% of NDEA FDR values.

Selection and cloning of variant-allele genomic regions. All genomic sequences were retrieved from the hg19 human genome assembly. For experimental testing, we selected variants with predicted DISs larger than 0 and included mutations near genes with evidence for ASD association, including those with coding LoF mutations (for example, *CACNA2D3*) and a proximal structural variant (for example, *SDC2*). We did not explicitly select mutations on the basis of proximity to TSSs and the chosen mutations lie from 7 bp to 324 kb away from the nearest TSS, with most variants lying farther than 5 kb from the nearest TSS (Supplementary Table 6). For each allele (sibling or proband), we either cloned 230 bp of genomic sequence amplified from proband lymphoblastoid cell lines or used FragmentGenes synthesized by Genewiz (Supplementary Table 6). In both cases, 15-nucleotide flanks on the 5′ and 3′ ends matched each flank of the plasmid cloning sites (Supplementary Table 6). Synthesized fragments were cut with KpnI and BglII and cloned into pGL4.23 (Promega) cut with the same enzymes. PCR-amplified genomic DNA was cloned into pGL4.23 blunt-end cut with EcoRV and Eco53kI using the GeneArtCloning method from Thermo Fisher Scientific. All constructs were verified by Sanger sequencing.

Luciferase reporter assays. Human neuroblastoma BE(2)-C cells were plated at 2×10^4 cells per well in 96-well plates and, 24 h later, were transfected with Lipofectamine 3000 (L3000-015, Thermo Fisher Scientific) together with 75 ng of Promega pGL4.23 firefly luciferase vector containing 230 bp of human genomic DNA from the loci of interest (Supplementary Table 6) and 4 ng of pNL3.1 NanoLuc (shrimp luciferase) plasmid, for normalization of transfection conditions. Forty-two hours after transfection, luminescence was detected with the Promega NanoGlo Dual-Luciferase assay system (N1630) and a BioTek Synergy plate reader. Four to six wells per variant were tested in each experiment. Variants were tested in at least two separate experiments. For each sequence tested, the ratio of firefly luminescence (ASD allele) to NanoLuc luminescence (transfection control) was calculated and then normalized to empty vector (pGL4.23 with no insert) on the same plate. Statistics were calculated from fold change over empty vector values from each experiment and results from multiple replication experiments were combined with Fisher's combined probability test. For presentation of the data, we normalized the fold change over empty vector value of the proband allele to that of the sibling allele.

***SMEK1* minigene assays.** To construct the *SMEK1* minigene, the genomic region was amplified as two separate regions and then cloned into the pSG5 vector; the first region contained the upstream exon plus approximately 1,400 bp of intron, and the second region contained the alternative exon and the downstream exon plus approximately 1,400 bp of intron (primers in Supplementary Table 6). The mutant minigene was constructed by assembling the PCR-amplified vector backbone with synthetic gBlocks (IDT DNA) carrying the desired single-base mutation (GRCh37 chromosome 14, g.91932755G>A in RNA). Minigenes (2 µg) were transfected into SH-SY5Y cells and cells were harvested 48 h after transfection for immunoblotting or reverse transcription with quantitative PCR following standard protocols. Three independent experiments were performed for statistical comparison.

Contribution of de novo mutations to ASD in the SSC. For LoF and missense coding mutations, we used annotations from Supplementary Table 1 of the SSC exome study[2]. Out of a total of 2,508 probands, 331 ASD probands had at least one LoF oding mutation and 1,182 probands had at least one missense mutation. We estimated the expected number of background occurrences in probands using occurrences in unaffected siblings adjusted by the overall proband/sibling ratio, resulting in 221.8 for LoF and 1,105.0 for missense mutations. The final estimated contribution was determined by the differential between observed and background occurrences (for example, for LoF mutations, 331 minus 221.8 divided by 2,508 probands gives an estimated contribution of 5.4%). For noncoding mutations, we observed 1,086 probands with mean DIS > 0 (mean of the average DNA DIS and average RNA DIS) in comparison to a background occurrence of 1,009 mutations per 1,781 individuals (unaffected siblings). The differential of 1,086 minus a background of 1,009 gives an estimated contribution of 4.3%.

Statistical analysis. All details of the statistical tests are specified in the associated text or figure legends. The NDEA test is described in detail in the Methods.

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

Data availability

ASD WGS data can be obtained from the Simons Foundation Autism Research Initiative (SFARI). All variant predicted scores have been made available as supplementary material and an interactive web interface is available at https://hb.flatironinstitute.org/asdbrowser/.

Code availability

The code used in this study is available from https://hb.flatironinstitute.org/asdbrowser/help.

References

40. Liu, Y., Li, B., Tan, R., Zhu, X. & Wang, Y. A gradient-boosting approach for filtering de novo mutations in parent–offspring trios. *Bioinformatics* **30**, 1830–1836 (2014).

41. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0 (2013).
42. Moore, M. J. et al. Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.* **9**, 263–293 (2014).
43. Darnell, J. C. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
44. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
45. 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).
46. Cotney, J. et al. The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. *Nat. Commun.* **6**, 6404 (2015).
47. 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).
48. Harrow, J. et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
49. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
50. Yan, Q. et al. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl Acad. Sci. USA* **112**, 3445–3450 (2015).
51. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
52. Mi, H. et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **45**, D183–D189 (2017).
53. Geifman, N., Monsonego, A. & Rubin, E. The neural/immune Gene Ontology: clipping the gene ontology for neurological and immunological systems. *BMC Bioinformatics* **11**, 458 (2010).
54. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **620**, 267–284 (2008).

nature research

Corresponding author(s): Olga Troyanskaya & Robert Darnell

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒ Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	no software was used.
Data analysis	DeepSEA 2.0 - for DNA variant prioritization Seqweaver 0.1 - for RNA variant prioritization xgboost 0.71 - for logistic regression model CTK package 1.0.5 - for CLIP processing GATK 3.8 - for variant calling DNMFilter 0.1 - for identifying de novo mutations R - 3.5.1 for statistically analysis including multiple hypothesis correction

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All ASD WGS data can be obtained from the Simons Foundation Autism Research Initiative (SFARI). All variant predicted disease impact scores have been made available as supplementary material.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size, all available samples were used from the Simons Simplex Collection.
Data exclusions	No data were excluded from the analyses.
Replication	Minimum of 2 biological replicates were conducted for the Luciferase Reporter Assays. All reported results are successfully replicated in all attempts.
Randomization	No systematic randomization was used.
Blinding	Our analysis were done using predefined Simons Simplex Collection families.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	BE(2)-C ATCC
Authentication	Low-passage cells purchased from ATCC were used and maintained as per ATCC.
Mycoplasma contamination	Negative by in-lab PCR assay
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.