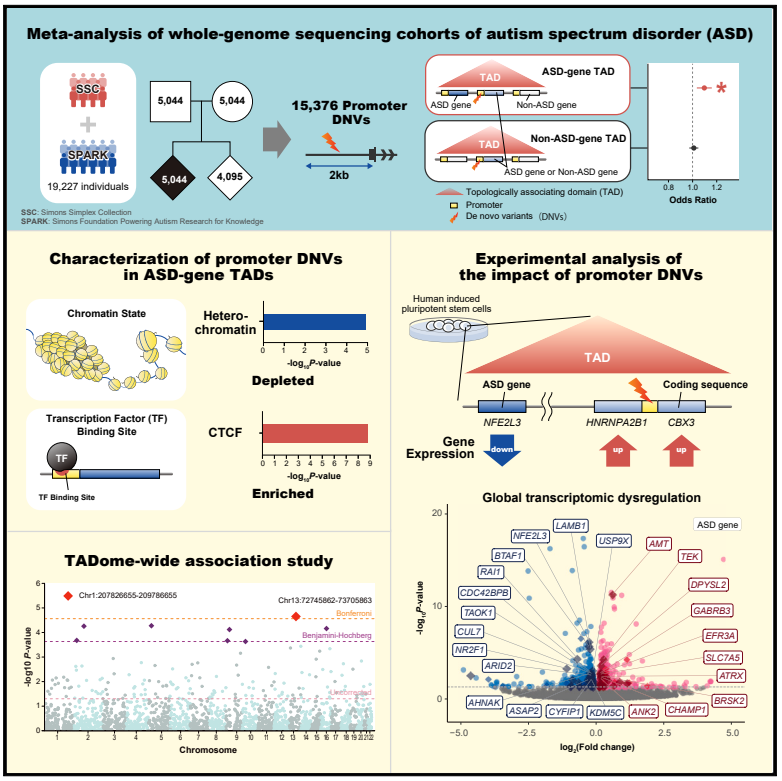
Cell Genomics

Article

Topologically associating domains define the impact of *de novo* promoter variants on autism spectrum disorder risk

Graphical abstract



Highlights

- Analysis of promoter DNVs in 5,044 ASD and 4,095 siblings by WGS using TAD information
- Specific association between ASD and promoter DNVs within TADs containing ASD genes
- Identification of TADs with enrichment of promoter DNVs in ASD
- Experimental validation of the effect of single promoter DNVs on multiple genes

Authors

Takumi Nakamura, Junko Ueda, Shota Mizuno, ..., Hirona Yamamoto, Tomonori Hara, Atsushi Takata

Correspondence

junko.hayashi.aa@riken.jp (J.U.), atsushi.takata@riken.jp (A.T.)

In brief

Nakamura, Ueda, and Mizuno et al. demonstrated that promoter *de novo* variants located in three-dimensional proximity to known autism spectrum disorder (ASD) genes are associated with disease risk. They experimentally showed that these variants influence not only the immediately downstream genes but also multiple distant genes, including those involved in ASD.

Nakamura et al., 2024, Cell Genomics *4*, 100488
February 14, 2024
https://doi.org/10.1016/j.xgen.2024.100488

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Topologically associating domains define the impact of *de novo* promoter variants on autism spectrum disorder risk

Takumi Nakamura,[1,5] Junko Ueda,[1,5,*] Shota Mizuno,[1,5] Kurara Honda,[1] An-a Kazuno,[1] Hirona Yamamoto,[1,2] Tomonori Hara,[1,3] and Atsushi Takata[1,4,6,*]

[1]Laboratory for Molecular Pathology of Psychiatric Disorders, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

[2]Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan

[3]Department of Organ Anatomy, Tohoku University Graduate School of Medicine, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan

[4]Research Institute for Diseases of Old Age, Juntendo University Graduate School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan

[5]These authors contributed equally

[6]Lead contact

*Correspondence: junko.hayashi.aa@riken.jp (J.U.), atsushi.takata@riken.jp (A.T.)

https://doi.org/10.1016/j.xgen.2024.100488

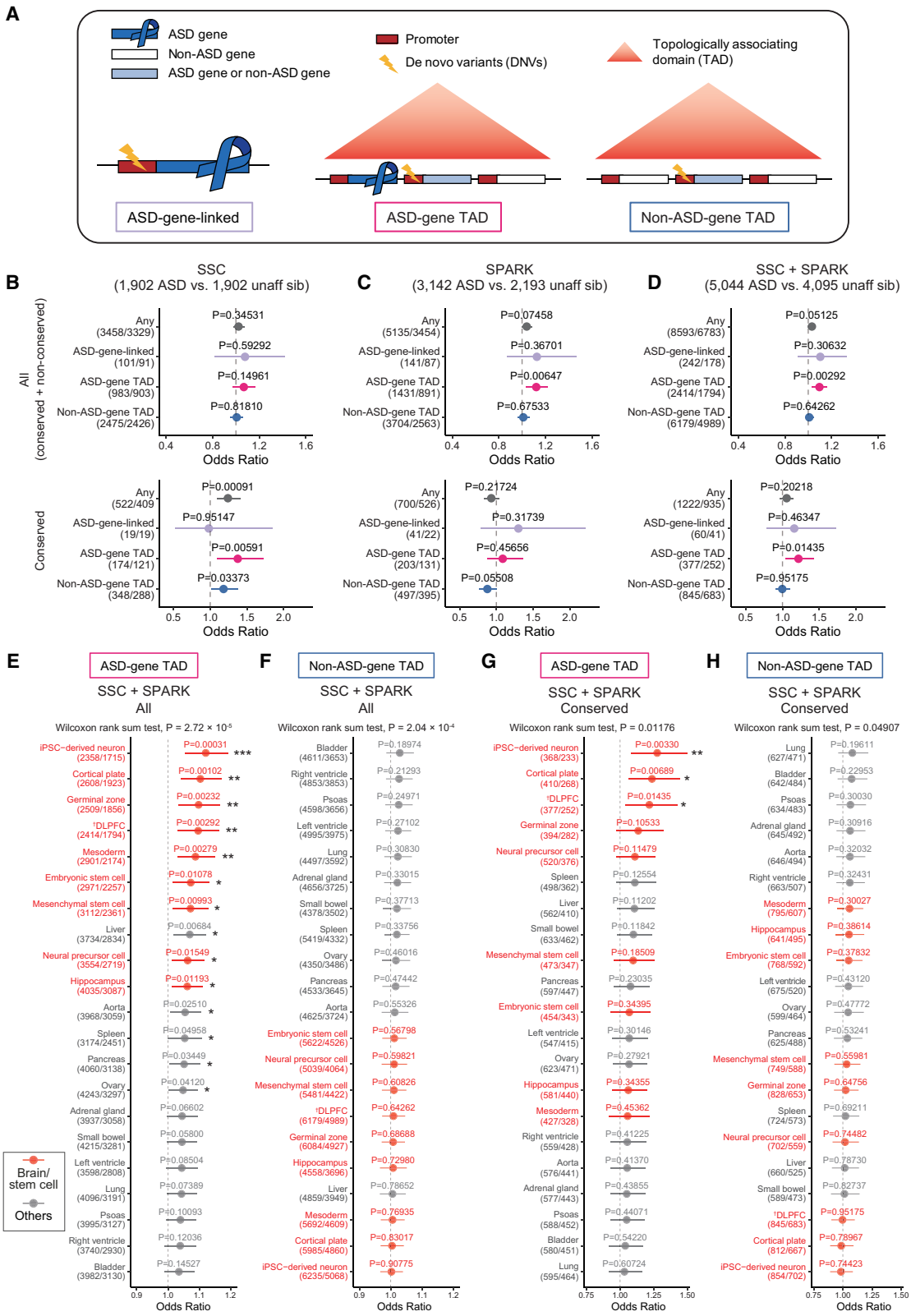
SUMMARY

Whole-genome sequencing (WGS) studies of autism spectrum disorder (ASD) have demonstrated the roles of rare promoter *de novo* variants (DNVs). However, most promoter DNVs in ASD are not located immediately upstream of known ASD genes. In this study analyzing WGS data of 5,044 ASD probands, 4,095 unaffected siblings, and their parents, we show that promoter DNVs within topologically associating domains (TADs) containing ASD genes are significantly and specifically associated with ASD. An analysis considering TADs as functional units identified specific TADs enriched for promoter DNVs in ASD and indicated that common variants in these regions also confer ASD heritability. Experimental validation using human induced pluripotent stem cells (iPSCs) showed that likely deleterious promoter DNVs in ASD can influence multiple genes within the same TAD, resulting in overall dysregulation of ASD-associated genes. These results highlight the importance of TADs and gene-regulatory mechanisms in better understanding the genetic architecture of ASD.

INTRODUCTION

Autism spectrum disorder (ASD) is a common neurodevelopmental condition characterized by impaired social communication and restricted, repetitive patterns of behavior seen in approximately 2% of children.[1] Studies have consistently demonstrated the highly heritable nature of ASD, which is contributed to by both common and rare variants. In particular, large-scale genome-wide association studies (GWASs) have shown that more than 10% of the liability can be explained by considering all common single-nucleotide polymorphisms (SNPs).[2–4] Another significant part of the liability can be attributed to rare single-nucleotide variants (SNVs), insertions/deletions (indels), and copy-number variants (CNVs) affecting protein-coding regions.[5–11] Some clinical studies reported that diagnostic gene-disruptive rare variants are observed in around 10% of individuals with ASD.[12–14] However, these common SNPs and rare protein-coding variants cannot explain a large part of the high heritability of ASD, which has been reported to be as high as 80%.[15] Therefore, it is presumed that rare non-coding variants contribute to at least some of the unresolved genetic architecture of ASD. To delineate the role of rare non-coding variants, several pioneering whole-genome sequencing (WGS) studies have been conducted, and one of the major findings of these studies is that promoter *de novo* variants (DNVs), especially those in evolutionarily conserved sequences, confer the risk of ASD.[16,17] The next important step should be to elucidate the mechanisms through which these DNVs confer the ASD risk and to identify their subsets associated with ASD with larger effect sizes.

To address this, here, we analyze two large-scale ASD WGS datasets from the Simons Simplex Collection[16] (SSC; 1,902 ASD probands and 1,902 unaffected siblings) and the Simons Foundation Powering Autism Research for Knowledge[18] (SPARK; 3,142 ASD probands and 2,193 unaffected siblings) and show that topologically associating domains (TADs), genomic regions where regulatory elements frequently interact with one another,[19] are informative when estimating the impact of promoter DNVs on the ASD risk. In particular, we observed that promoter DNVs within TADs containing known ASD genes, but not those in the other non-ASD-gene TADs, are significantly associated with the case-control labels. Subsequently, we



(legend on next page)

characterized gene-regulatory features of promoter DNVs in ASD probands located within TADs containing ASD genes, analyzed specific TADs statistically enriched for promoter DNVs in ASD, and experimentally investigated the impact of selected promoter DNVs on other genes in the same TAD and overall transcriptomic profiles.

RESULTS

Association analysis of promoter DNVs with ASD in light of TADs and their gene contents

As described above, previous studies have reported an association of promoter DNVs with ASD.[16,17] Considering the mechanism through which promoter DNVs contribute to the pathogenesis of ASD, it can be easily assumed that these variants may dysregulate the expression of their immediately downstream genes. If so, promoter DNVs upstream of known ASD genes should be particularly enriched in ASD probands. However, by reanalyzing the data of promoter DNVs in a large-scale WGS study of 1,902 ASD quads (proband, unaffected sibling, and parents) from SSC,[16] we found no significant association of such ASD-gene-linked promoter DNVs with ASD (Figure 1A, left) in ASD probands (Figure 1B, $p = 0.593$, odds ratio [OR] = 1.078, for all DNVs and $p = 0.951$, OR = 0.981, for conserved DNVs; note that we applied logistic regression tests with the ages of the father and mother at the birth of the child as covariates because these are known factors influencing the DNV rates[20,21]; ASD genes were defined by the SFARI gene database[22]; see STAR Methods for details). Following this observation, we investigated an alternative hypothesis. Recently, it has been demonstrated that disruption of a promoter not only reduces the expression of the downstream gene but also induces the retargeting of an enhancer originally bound to that promoter to other promoters in the same TAD and alters the expression of genes downstream of the retargeted promoters.[23] Being inspired by this newly reported phenomenon, we speculated that the presence of an ASD gene(s) within the same TAD, instead of the property of the immediately downstream gene, might be associated with the pathogenicity of promoter DNVs.

To test this hypothesis, we performed an analysis of the SSC dataset by classifying promoter DNVs into those within TADs containing known ASD genes among genes other than that immediately downstream of the DNV (referred to as ASD-gene TADs) and the other non-ASD-gene TADs (Figure 1A, middle and right). In this dataset, there are 3,458 and 3,329 DNVs with promoter annotation in 1,902 ASD probands and 1,902 unaffected siblings, respectively (Tables S1 and S2). The brain TADs were defined by using the Hi-C (high-throughput chromosome conformation capture) data of adult human dorsolateral prefrontal cortex (DLPFC) generated through the PsychENCODE project.[24] When we analyzed the association of these three types of promoter DNVs with the case-control labels (see STAR Methods for details), we observed that ASD-gene TAD promoter DNVs consistently showed higher ORs than the other non-ASD-gene TAD promoter DNVs (Figure 1B; Table S3A). To further evaluate whether ASD-gene TAD promoter DNVs contribute more to the disease risk than the other promoter DNVs, we subsequently analyzed the SPARK dataset (3,142 ASD probands and 2,193 unaffected siblings; Table S1). To ensure consistency in terms of the analytical pipeline across cohorts, we followed the methods for DNV calling used in the study of SSC by An et al.[16] and identified comparable numbers of DNVs in the SPARK dataset (Figure S1; Table S1, per-individual DNV counts [average ± SD] = 59.5 ± 13.0 in ASD probands and 59.7 ± 13.1 in unaffected siblings). Before analyzing promoter DNVs, we confirmed that loss-of-function (LoF, e.g., nonsense, frameshift, and splice site) DNVs, especially those in constraint genes with the first and second deciles of LoF observed/expected upper-bound fraction (LOEUF) scores,[25] are significantly overrepresented in ASD in SPARK (Figure S2). On the other hand, there was no significant case-control difference in the numbers of synonymous DNVs (Figure S2). These predictable results support the overall validity of the DNV dataset in SPARK generated in this study. In the analysis of promoter DNVs not considering TADs, we identified 5,135 and 3,454 promoter DNVs in ASD probands and unaffected siblings, respectively (the full list is available in Table S2), and found a trend toward association of promoter DNVs with ASD (Figure 1C; Table S3A, $p = 0.0746$, OR = 1.040). On the other hand, there was no significant association between promoter DNVs in conserved regions and ASD (Figure 1C; Table S3A, $p = 0.217$, OR = 0.931). We then performed an analysis of the promoter DNVs in the SPARK dataset by classifying these into ASD-gene-linked, ASD-gene TAD, and non-ASD-gene TAD DNVs

Figure 1. Association analysis of promoter DNVs in ASD in light of TADs and their gene contents

(A) Schematics of ASD-gene-linked (left), ASD-gene TAD (middle), and non-ASD-gene TAD (right) promoter DNVs. ASD-gene-linked promoter DNVs are those immediately upstream of known ASD genes. ASD-gene TAD promoter DNVs are those within TADs containing known ASD genes among genes other than that immediately downstream of the DNV. Non-ASD-gene TAD promoter DNVs are those within TADs not containing known ASD genes among genes other than that immediately downstream of the DNV. Note that the gene immediately downstream of the promoter hit by a DNV can be either an ASD or a non-ASD gene for ASD-gene TAD or non-ASD-gene TAD promoter DNVs.

(B–D) Analysis of the association between ASD and any, ASD-gene-linked, ASD-gene TAD, or non-ASD-gene TAD promoter DNVs, which are defined based on the DLPFC TADs. Results of the analyses in SSC (B) and SPARK (C) and the meta-analysis of these two cohorts (D) for all (top) and evolutionarily conserved (bottom) promoter DNVs are shown with the uncorrected p values. The ORs and p values were calculated by logistic regression analysis using the ages of the father and mother at the birth of the child as covariates in the SSC and SPARK datasets and the inverse-variance weighted method in the meta-analysis.

(E–H) Analysis of the association between ASD and ASD-gene TAD (E and G) or non-ASD-gene TAD (F and H) promoter DNVs using TAD datasets derived from various tissues and cell lines. Results of the meta-analyses of SSC and SPARK cohorts for all (E and F) and evolutionarily conserved (G and H) promoter DNVs are shown with the uncorrected p values. Results are sorted according to the ORs, and the statistical differences in ORs between the brain and stem cells and the other tissues calculated by the two-tailed Wilcoxon rank-sum test are displayed at the top of each chart. *uncorrected $p < 0.05$, **Benjamini-Hochberg-corrected $p < 0.05$, ***Bonferroni-corrected $p < 0.05$ (based on the 84 tests in total). †Same as the DLPFC dataset used in (B)–(D). In (B)–(H), the error bars indicate 95% confidence intervals, and the numbers of DNVs in ASD probands/unaffected siblings included in each analysis are shown in parentheses.

(Figure 1C). Again, we observed higher ORs for ASD-gene TAD promoter DNVs, with a significant association between all (not restricted to conserved) ASD-gene TAD promoter DNVs and ASD (Figure 1C; Table S3A). Regarding ASD-gene-linked promoter DNVs, we did not observe their significant association with ASD (Figure 1C). Meanwhile, the observed OR for them in the analysis of all promoter DNVs was similar to that for ASD-gene TAD promoter DNVs (Figure 1C, top) and rather higher in the analysis of conserved promoter DNVs (Figure 1C, bottom). By meta-analyzing the results in the SSC and SPARK datasets, we observed a significant association between ASD-gene TAD promoter DNVs and ASD (Figure 1D; Table S3A, $p = 0.00292$, OR = 1.096, for all DNVs and $p = 0.0144$, OR = 1.217, for conserved DNVs, inverse-variance weighted method), whereas non-ASD-gene TAD promoter DNVs showed no association, with ORs close to 1 (Figure 1D; Table S3A, $p = 0.643$, OR = 1.009, for all DNVs and $p = 0.952$, OR = 0.997, for conserved DNVs).

To further evaluate the reproducibility of the above findings and to investigate the tissue specificity, we then repeatedly performed a meta-analysis of SSC and SPARK cohorts using TAD datasets derived from various tissues and stem cell lines (Figures 1E–1H; Table S3B, see STAR Methods for details). In the analyses of all (conserved and non-conserved) ASD-gene TAD promoter DNVs, we observed that these DNVs in any of the developing and adult brain tissues and cultured stem cells were significantly associated with ASD at an uncorrected $p < 0.05$ (Figure 1E; Table S3B). In contrast, there was no or only modestly significant association when TADs in peripheral tissues were used, with a statistically significant difference in ORs between brain and stem cells and peripheral tissues (Figure 1E, $p = 2.72 \times 10^{-5}$, two-tailed Wilcoxon rank-sum test). Similar to the analysis using the data of TADs in DLPFC, in all tissues and cell lines, non-ASD-gene TAD promoter DNVs were not significantly associated with ASD (Figure 1F). The ORs observed in brain and stem cells were significantly smaller than those in peripheral tissues (Figure 1F, $p = 2.04 \times 10^{-4}$). These patterns of associations were largely consistent when we focused on promoter DNVs hitting conserved regions, again with significant differences in ORs between brain and stem cells and peripheral tissues (Figures 1G and 1H; Table S3B). Among individual analyses of various tissues and cell lines, ASD-gene TAD promoter DNVs in human induced pluripotent stem cell (iPSC)-derived neurons exhibited the most significant association with ASD, which remained significant after Bonferroni correction was performed with the total number of tests (Figure 1E, Bonferroni-corrected $p = 0.0258$, OR = 1.121, n = 84 tests = 21 tissues × 4 promoter DNV types [all and conserved; ASD-gene TAD and non-ASD-gene TAD]). Significant associations after the Benjamini-Hochberg correction were also found in all ASD-gene TAD promoter DNVs in the cortical plate, germinal zone, DLPFC, and mesoderm (Figure 1E) and conserved ASD-gene TAD promoter DNVs in iPSC-derived neurons (Figure 1G). These results collectively indicate that the gene contents of TADs can influence the pathogenic relevance of promoter DNVs, and in particular, the presence of an ASD gene(s) within the same TAD in the brain and in early developmental cells is associated with the impact of promoter DNVs on ASD risk.

Characterization of promoter DNVs in ASD-gene TADs

We next sought to characterize the properties of ASD-gene TAD promoter DNVs in ASD probands using various features such as epigenetic state of the variant sites and the length and contents of TADs. In this analysis, we used the list of ASD-gene TAD promoter DNVs defined in our initial analysis based on the DLPFC TADs, considering that this dataset of TADs distributed through the PsychENCODE project is one of the most widely used resources in studies of neuropsychiatric disorders. First, we annotated the positions hit by promoter DNVs with predicted chromatin state using the data of 15 chromatin states across 17 tissues/cells from the Roadmap Epigenomics Project[26] and then tested if the proportion of ASD-gene TAD promoter DNVs hitting regions with each chromatin state was significantly different between ASD probands and unaffected siblings (see STAR Methods for details). In the analysis of individual tissues and chromatin states, there was no significant annotation category after multiple testing correction with the total number of tests (Figure 2A; Table S4, the smallest uncorrected $p = 0.00119$, two-tailed Fisher's exact test, n = 237 tests [17 tissue/cell groups × 15 chromatin states − 18 unavailable annotations]). On the other hand, when we analyzed the results collectively by each chromatin state by calculating synthetic p values taking the non-independence of each tissue/cell into account (see STAR Methods for details), we found that ASD-gene TAD promoter DNVs in ASD probands are significantly depleted in heterochromatin regions (9_Het; combined $p = 1.21 \times 10^{-5}$, Fisher's method). Although there was no significance considering the number of chromatin states tested (n = 15), we also observed that ASD-gene TAD promoter DNVs in other repressive regions (e.g., 13_ReprPC, repressed PolyComb, and 14_ReprPCWk, weak repressed PolyComb) are less frequent (OR < 1 in the majority of tissue/cell groups) in ASD probands, whereas these variants in transcribed regions (e.g., 3_TxFlnk, transcription at gene 5′ and 3′, and 5_TxWk, weak transcription) are relatively frequent in ASD probands (Figure 2A; Table S4). There was no obvious tissue specificity in these patterns of depletion/enrichment, and the brain tissues showed typical results, such as depletion of ASD-gene TAD promoter DNVs in transcriptionally repressive regions in ASD probands.

We next evaluated whether ASD-gene TAD promoter DNVs in ASD probands were enriched in regions bound to specific transcription factors (TFs) in the neural tissues/cells using ChIP-Atlas,[27] a comprehensive database of chromatin immunoprecipitation sequencing (ChIP-seq) and other experiments for gene-regulatory elements. After the data were filtered based on the numbers of DNVs within TF-bound regions and the available experiments, 148 experiments for 13 TFs remained (see STAR Methods for details). When we analyzed individual experiments, no dataset showed statistical significance after conservative multiple testing correction with the total number of tests (Figure S3A; Table S5, the smallest uncorrected $p = 0.00463$, two-tailed Fisher's exact test, n = 148 tests). On the other hand, analysis of each TF showed that ASD-gene TAD promoter DNVs in ASD probands were significantly enriched in regions bound to CTCF, a protein critical for the formation of TAD boundaries and also known to bind to promoter regions to engage in enhancer-promoter interactions[23] (Figure 2B,

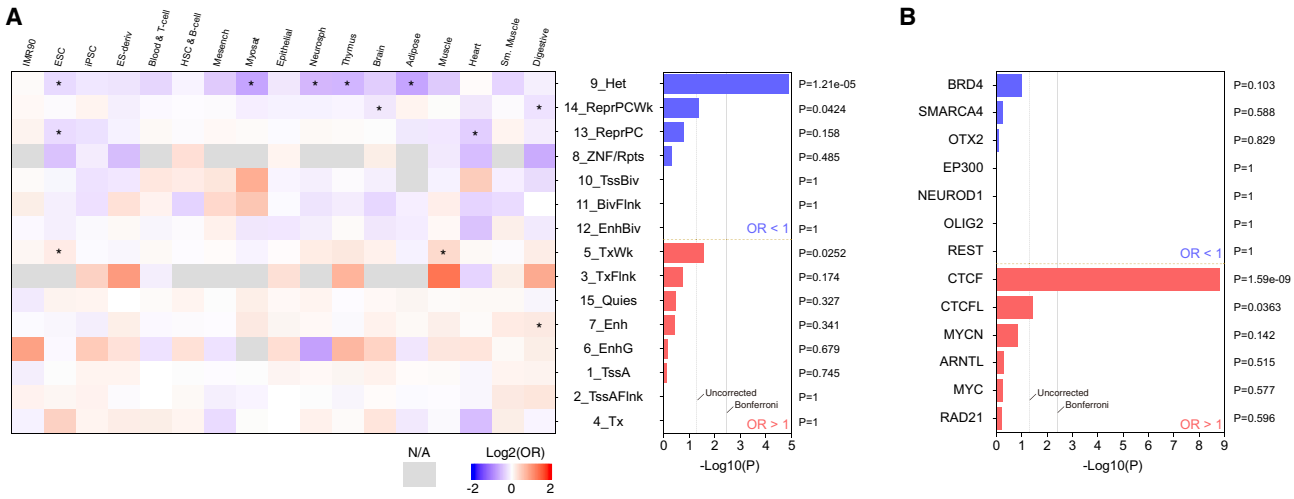


Figure 2. Characterization of promoter DNVs in ASD-gene TADs

(A) Left: heatmap representation of the chromatin states (ChromHMM core 15-state model) of the regions hit by ASD-gene TAD promoter DNVs in ASD probands and unaffected siblings. The data of various tissues/cell lines from the Roadmap Epigenomics Project[26] were used. The heatmap is color coded by the ORs calculated from the numbers of ASD-gene TAD promoter DNVs in ASD probands and unaffected siblings hitting and not hitting genomic regions with the corresponding chromatin state (see STAR Methods for details). Red indicates that the regions with the corresponding chromatin state are more likely to be hit by ASD-gene TAD promoter DNVs in ASD probands than those in unaffected siblings. When the total number of ASD TAD promoter DNVs in a given chromatin state of a given tissue/cell was less than 10, such combinations were excluded from the statistical analysis and are indicated in light gray in the heatmap. The asterisks in the heatmap indicate uncorrected $p < 0.05$ (two-tailed Fisher's exact test). Right: bar plot of synthetic p values calculated from the p values for each tissue/cell type (see STAR Methods and Table S4 for details).

(B) Bar plot representing depletion or enrichment of ASD-gene TAD promoter DNVs in ASD probands in the genomic regions bound to each TF. As in (A), the p values for each TF and each neural tissue/cell were calculated by two-tailed Fisher's exact test and then synthesized per TF using Fisher's methods (see STAR Methods, Table S5, and Figure S3 for details). In (A) and (B), the blue and red bars indicate OR < 1 (depletion) and OR > 1 (enrichment) in ASD, respectively. Het, heterochromatin; ReprPCWk, weak repressed PolyComb; ReprPC, repressed PolyComb; ZNF/Rpts, zinc finger genes and repeats; TssBiv, bivalent/poised TSS; BivFlnk, flanking bivalent TSS/Enh; EnhBiv, bivalent enhancer; TxWk, weak transcription; TxFlnk, transcription at gene 5′ and 3′; Quies, quiescent/low; Enh, enhancers; EnhG, genic enhancers; TssA, active TSS; TssAFlnk, flanking active TSS; Tx, strong transcription.

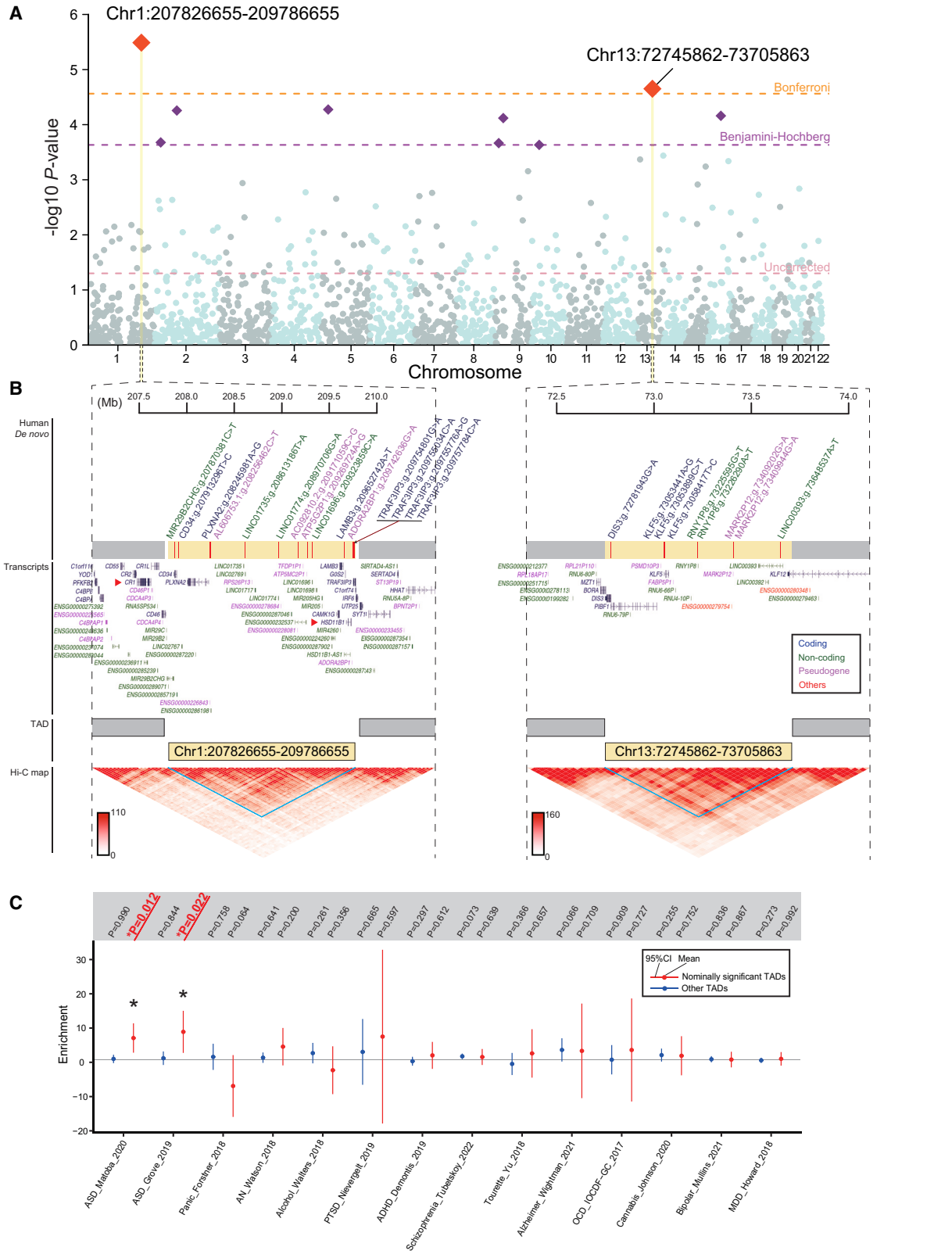
combined $p = 1.59 \times 10^{-9}$, Fisher's method), and its paralog CTCFL with nominal significance (Figure 2B, $p = 0.0363$).

When we evaluated if the pathogenicity scores for non-coding variants predicted by various computational tools, such as CADD,[28,29] Eigen,[30] and LINSIGHT[31] (annotated by regBase,[32] see STAR Methods for details), differed between ASD-gene TAD promoter DNVs in ASD probands and unaffected siblings, no significant differences were observed in any of the analyses using 22 scoring systems (Figure S3B, two-tailed Wilcoxon rank-sum test). Meanwhile, in 20 of the 22 analyses, except for those with CDTS[33] and GenoCanyon,[34] the average Phred-scaled scores for ASD-gene TAD promoter DNVs in ASD probands were higher (predicted to be more deleterious) than in unaffected siblings. This may suggest that these DNVs in ASD are on average more deleterious, but accurate discrimination using existing scoring tools is difficult.

We also annotated each ASD-gene TAD promoter DNV with information related to the TAD containing them, such as TAD length, the numbers of genes and regulatory elements contained, and the distance between the DNV and the nearest ASD gene, and analyzed case-control differences. We found that the distribution of distances between the nearest ASD gene and the ASD-gene TAD promoter DNV was biased toward shorter in ASD probands compared with unaffected siblings (Figure S3C, $p = 0.0353$, one-tailed two-sample Kolmogorov-Smirnov test). On the other hand, no such biases were observed in the distribution of the length of TADs containing ASD-gene TAD promoter DNVs (Figure S3C, $p = 0.975$) or the distance between the DNV and the nearest TAD boundary (Figure S3C, $p = 0.470$). Regarding the numbers of genic transcription start sites (TSSs) and brain enhancers (defined by using the PsychENCODE resource[24]) in TADs containing ASD-gene TAD promoter DNVs, we found no significant differences between ASD probands and unaffected siblings, while there was a trend for TADs encompassing ASD-gene TAD promoter DNVs observed in ASD probands to harbor more enhancers (Figure S4, $p = 0.0842$, two-tailed Wilcoxon rank-sum test).

TADome-wide association study

TADs, where regulatory elements interact with one another frequently, can be considered functional units, and there might be specific TADs significantly enriched for promoter DNVs in ASD. We next analyzed this in a comprehensive, i.e., TADome-wide, manner. To this end, we calculated the expected number of *de novo* SNVs in the promoter regions of each TAD using an established trinucleotide-context mutational model[35] and compared this with the observed number of promoter *de novo* SNVs (see STAR Methods for details). By performing a TADome-wide association study (TADWAS) of 1,826 TADs with one or more promoter *de novo* SNVs in ASD, we identified two TADs, each on



(legend on next page)

chromosome 1 or 13, that achieved statistical significance after Bonferroni correction was performed with the number of TADs (Figures 3A and 3B; Table S6, threshold $p = 0.05/1,826 = 2.74 \times 10^{-5}$). Also, 9 and 156 TADs, respectively, had Benjamini-Hochberg-corrected and uncorrected $p < 0.05$.

Of the two significant TADs after Bonferroni correction, 15 promoter *de novo* SNVs in the TAD spanning chr1:207826655–209786655 (hg38) were identified in 5,044 ASD probands. This observation is unlikely to have occurred by chance, given that the expected number is 3.42 ($p = 3.24 \times 10^{-6}$, Poisson exact test). This TAD contains two SFARI category 3 ASD-associated genes, *CR1* encoding complement C3b/C4b receptor 1 and *HSD11B1* encoding hydroxysteroid 11-β dehydrogenase 1. Of the 15 promoter *de novo* SNVs, 4 are located upstream of *TRAF3IP3*. Based on the theoretical mutation rate for the *TRAF3IP3* promoter region, the probability of observing 4 or more *de novo* SNVs was 1.23×10^{-6}, suggesting the pathogenic relevance of DNVs hitting this specific promoter. The other significant TAD spanning chr13:72745862–73705863 encompasses 9 promoter *de novo* SNVs in ASD, whereas the expected number is 1.46 ($p = 2.23 \times 10^{-5}$, Poisson exact test). Because this TAD does not contain any SFARI ASD genes, there might be unknown disease-associated genes, possibly including non-coding ones. Snapshots of the mapped sequence reads supporting the promoter *de novo* SNVs in these TADs on chromosomes 1 and 13 are shown in Figures S5–S12.

Next, we examined if the heritability of various neuropsychiatric phenotypes is enriched in the promoter regions of 156 TADs that showed nominal significance (uncorrected $p < 0.05$) in the TADWAS by linkage disequilibrium score regression (LDSC; see STAR Methods for details).[37,38] We observed that ASD heritability is significantly enriched in these regions, which was supported by both of the two large-scale ASD GWASs[3,4] (Figure 3C; Table S7). On the other hand, the heritability of other phenotypes[40–51] was not significantly enriched in the promoter regions of nominally significant TADs. In the promoters of the other non-significant TADs, there was no significant heritability enrichment for any neuropsychiatric phenotypes (Figure 3C). We also analyzed GWAS data from the UK Biobank for various traits[52] and found that none of them show stronger heritability enrichment in the promoter regions of nominally significant TADs than ASD (Figure S13; Table S7). These results provide convergent evidence that both rare *de novo* and common variants in the promoter regions of TADs with DNV enrichment in ASD contribute to the genetic predisposition for ASD. Possibly reflecting this, we found that the SNP (rs716219, association $p = 6.42 \times 10^{-9}$) that showed the most significant association with ASD in a GWAS meta-analysis[4] is located within a TAD (chr1:95254444–96894444) with nominally significant enrichment of promoter DNVs (Figure S14; Table S8, enrichment $p = 0.00713$). In addition, suggestive association signals ($p < 5 \times 10^{-5}$) in this GWAS meta-analysis were detected in three of the nine TADs with Benjamini-Hochberg-corrected $p < 0.05$ in the TADWAS (chr5:24199891–26719891, chr9:29240002–32360002, and chr9:12680000–15160002; Figures S15–S17; Table S8).

Experimental analysis of the impact of selected ASD TAD promoter DNVs using the iPSC model

The above results collectively provide statistical evidence that ASD TAD promoter DNVs, particularly those in the brain and stem cells, contribute to the genetic risk of ASD. We subsequently selected two of the ASD TAD promoter DNVs identified in SSC probands and experimentally evaluated their impact on the transcriptomic profile by introducing a DNV into human iPSCs using the CRISPR-Cas9 system (STAR Methods and Tables S9, S10, and S11).

The first ASD TAD promoter DNV, the chr7:g.26200781C>T variant, was selected based on the following criteria: (1) the position of the DNV is conserved across species; (2) the position of the DNV is shown to be bound to CTCF in neural cells, referring to the results in Figure 2B; (3) the TAD containing the DNV includes one or more high-confidence ASD genes (SFARI genes in categories S, 1, and 2) among the genes other than the gene immediately downstream of the promoter hit by the DNV; and (4) a suitable single-guide RNA (sgRNA) to introduce the variant can be designed (Figures 4A and 4B). The chr7:g.26200781C>T variant is located upstream of both the *HNRNPA2B1* and the *CBX3* genes. The TAD containing this DNV includes an SFARI category 2 gene *NFE2L3*, encoding NFE2-like BZIP transcription factor 3. After generating the iPSC lines with or without the chr7:g.26200781C>T variant using repair templates with the variant or reference allele (see STAR Methods for details), we performed RNA-sequencing (RNA-seq) analysis to identify genes differentially expressed between the knockin (three clones from one line) and the control (six clones from two lines) cells. We primarily investigated cells with a homozygous variant, anticipating that the effects of the DNV could be more clearly detected, whereas we also analyzed heterozygous knockin cell lines and evaluated the overall similarity between homozygous and heterozygous mutants (Figure S18). We also confirmed that there is no off-target variant at candidate sites in all cell lines used for the experiments.

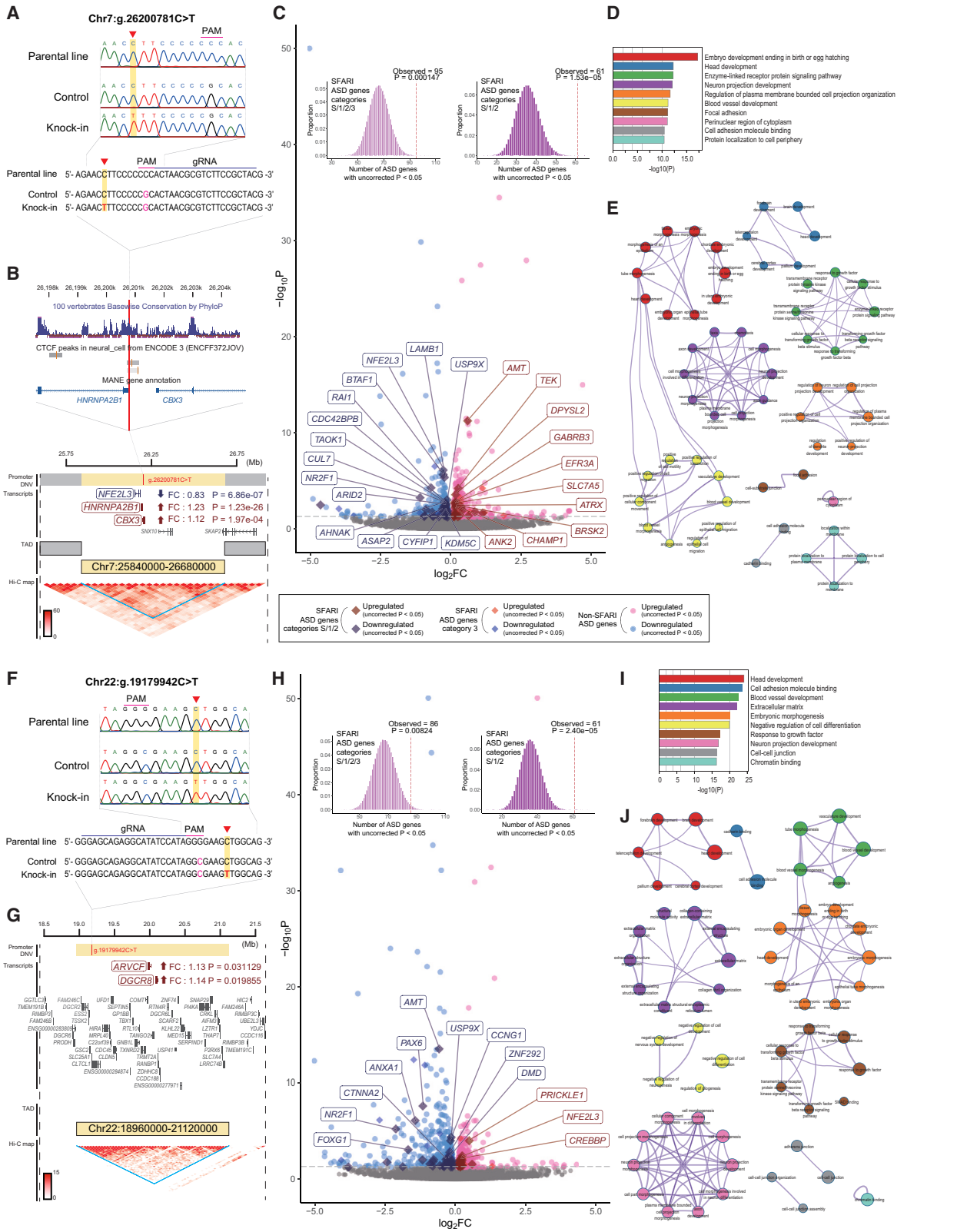
Within the TAD containing the chr7:g.26200781C>T variant (chr7:25840000–26680000), we observed significant upregulation of *HNRNPA2B1* (uncorrected $p = 1.23 \times 10^{-26}$; fold change [FC] = 1.32) and *CBX3* (uncorrected $p = 0.000197$; FC = 1.12),

Figure 3. TADome-wide association study and heritability enrichment analysis

(A) A Manhattan plot of the result of the TADWAS. TADs with Bonferroni-corrected $p < 0.05$ are labeled with the location of the TAD.

(B) Schematic representation of two TADs with TADome-wide significant enrichment of promoter DNVs in ASD. The positions of promoter DNVs are plotted onto the Hi-C interaction heatmap generated by the 3D Genome Browser[36] (tissue, Cortex_DLPFC; resolution, 40 kb). The red arrowheads indicate SFARI ASD genes.

(C) Heritability enrichment analysis of the promoter regions of 156 TADs with nominal significance (uncorrected $p < 0.05$) in the TADWAS (red) and the promoter regions of the other TADs (blue), which was performed by linkage disequilibrium score regression (LDSC)[37,38] with GWAS summary statistics obtained from studies by the Psychiatric Genomics Consortium (PGC)[39] and Matoba et al.[4] Enrichment was calculated from the proportions of SNPs in the regions of interest and the heritability explained by these SNPs. Results are sorted in the order of enrichment p values in the analysis of promoters in nominally significant TADs. The error bars indicate 95% confidence intervals.



(legend on next page)

the two genes downstream of the introduced DNV (Figure 4B). Intriguingly, *NFE2L3*, an ASD-associated gene[54] in this TAD, was significantly downregulated in the knockin cells, suggesting alteration of promoter-enhancer interactions within this TAD (uncorrected $p = 6.86 \times 10^{-7}$; FC = 0.83, Figure 4B). When we evaluated the overall effects of the chr7:g.26200781C>T variant on the transcriptomic profile, we identified a total of 1,208 differentially expressed genes (DEGs) at the level of uncorrected $p < 0.05$ in the knockin cells (Figure 4C; Table S12). By analyzing the properties of these DEGs, we found that these are significantly enriched for known ASD-associated genes (SFARI genes in categories S/1/2/3, Figure 4C insets, hypergeometric $p = 0.000147$), especially for high-confidence ASD genes (SFARI genes in categories S/1/2, Figure 4C insets, hypergeometric $p = 1.53 \times 10^{-5}$). Gene ontology (GO) enrichment analysis showed an overrepresentation of genes related to the development of the embryo and nervous system among these DEGs (Figures 4D and 4E; Table S13). Enrichment of terms involved in development and brain function, such as the overrepresentation of GABA receptor genes in the upregulated DEGs, was confirmed in an analysis using another statistical approach, gene set enrichment analysis (GSEA) (Figures S19A and S19B; Table S14). Therefore, this DNV appears to cause overexpression of the downstream genes, rather than inducing promoter disruption and enhancer retargeting, and as the consequence, *NFE2L3* within the same TAD and other genes associated with ASD and neurodevelopment are dysregulated in the knockin cells.

The second chr22:g.19179942C>T variant was chosen primarily based on its location. Given the expected consequence of enhancer release and retargeting, it is possible that disruption of a single promoter by a DNV in a region where genomic duplication is known to be associated with the risk of ASD, such as 15q11–q13, 16p11.2, and 22q11.2, may cause upregulation of multiple genes at the corresponding loci and thereby contribute to disease risk. We explored ASD TAD promoter DNVs in these regions and selected the chr22:g.19179942C>T variant hitting an evolutionarily conserved region for which we could design a suitable sgRNA. After introducing this variant into iPSCs (Figure 4F), we performed RNA-seq to identify genes differentially expressed between the knockin (three clones from one line) and the control (six clones from two lines) iPSCs.

In the TAD spanning chr22:18960000–21120000 containing the chr22:g.19179942C>T variant, we observed upregulation of two genes, *ARVCF* (uncorrected $p = 0.0311$, FC = 1.13) and *DGCR8* (uncorrected $p = 0.0199$; FC = 1.14), in the knockin cells at the level of uncorrected $p < 0.05$ (Figure 4G). Although these are not significant after correcting for the number of protein-coding genes in this TAD subjected to the DEG analysis (n = 41), nominally significant upregulation of *DGCR8* is potentially interesting, because this gene is considered a key gene that explains the psychiatric symptoms observed in the 22q11.2 copy number variation syndromes.[55–57] Among the 40 genes excluding *SLC25A1* immediately downstream of the chr22:g.19179942C>T variant, 29 showed an increase in their expression (FC > 1, Figure 4G). This significant bias toward upregulation ($p = 0.00643$, two-tailed binomial test) suggests that the chr22:g.19179942C>T variant induces overall transcriptional activation in this TAD. Regarding the transcriptomic alterations in iPSCs with the chr22:g.19179942C>T variant, similar to the chr7:g.26200781C>T mutants, we observed a significant overrepresentation of known ASD-associated genes among the DEGs (Figure 4H insets, hypergeometric $p = 0.00824$ for SFARI genes in categories S/1/2/3 and $p = 2.40 \times 10^{-5}$ for SFARI genes in categories S/1/2, n = 1,226 DEGs at uncorrected $p < 0.05$; Table S15). GO analysis of DEGs showed enrichment of genes involved in brain and blood vessel development, cell adhesion, and others (Figures 4I and 4J; Table S16). In the GSEA, we found that many terms are enriched in the downregulated DEGs, which include those related

Figure 4. Local and global transcriptomic dysregulation in iPSCs with selected promoter DNVs

(A) Sanger sequencing of the region surrounding chr7:26200781 in the parental iPSC line and the control and knockin lines. Note that a variant at the protospacer adjacent motif (PAM) sequence was introduced into both knockin and control lines to avoid the generation of unwanted variants due to repeated cleavage of the target site.

(B) Top: magnified view of the region proximal to the chr7:g.26200781C>T variant (the red vertical line) with information on the 100 vertebrate PhyloP conservation scores, CTCF peaks in neuronal cells, and the Matched Annotation from NCBI and EMBL-EBI (MANE)[53] gene annotation. Bottom: genes in the TAD containing the chr7:g.26200781C>T variant (chr7:25840000–26680000) and the Hi-C interaction heatmap of this TAD (the cyan triangle; see STAR Methods for details). In the gene track, differentially expressed genes within this TAD (*NFE2L3*, *HNRNPA2B1*, and *CBX3*) are highlighted along with the fold change (FC) and uncorrected p values.

(C) A volcano plot showing differentially expressed genes (DEGs) in cells with the chr7:g.26200781C>T variant. The overlap between DEGs with uncorrected $p < 0.05$ and SFARI ASD genes (left, categories S/1/2/3 [a broader list]; right, categories S/1/2 [a list of high-confidence genes]) is shown in the insets, where the histograms show the hypergeometric distribution of the expected counts and the red vertical dotted lines indicate the observed count along with the corresponding p value. In the volcano plot, SFARI S/1/2 genes with Benjamini-Hochberg-corrected $p < 0.3$ (uncorrected $p < \sim 0.0075$) are labeled.

(D) The top 10 Gene Ontology (GO) terms enriched among DEGs in chr7:g.26200781C>T iPSCs at uncorrected $p < 0.05$.

(E) Network visualization of the enriched GO terms in (D). Node sizes are proportional to the enrichment p values. The cluster colors correspond to those in the bar plot in (D).

(F) Sanger sequencing of the region surrounding chr22:19179942 in the parental, control, and knockin lines.

(G) Genes in the TAD containing the chr22:g.19179942C>T variant (chr22:18960000–21120000) and the Hi-C interaction heatmap. Two genes with uncorrected $p < 0.05$ (*ARVCF* and *DGCR8*) are highlighted.

(H) A volcano plot showing DEGs in iPSCs with the chr22:g.19179942C>T variant and the inset histograms indicating the overlap between DEGs with uncorrected $p < 0.05$ and SFARI ASD genes. SFARI S/1/2 genes with Benjamini-Hochberg-corrected $p < 0.3$ (uncorrected $p < \sim 0.0081$) are labeled.

(I) The top 10 GO terms enriched among DEGs in chr22:g.19179942C>T iPSCs at uncorrected $p < 0.05$.

(J) Network visualization of the enriched GO terms in (I). In (C) and (H), genes with $p < 1 \times 10^{-50}$ or $|\log2\ FC| > 5$ (n = 1 and 3 in C and H, respectively) are shown at the edge of the plot for visibility.

to the extracellular matrix, morphogenesis, and forebrain generation (Figures S19C and S19D; Table S17).

To further characterize the DEGs in these iPSC lines with the promoter DNV of interest, we performed a comparison between the DEGs in our iPSC models and the DEGs reported in the currently largest ASD postmortem brain study that analyzed a total of 725 brain samples across 11 cortical regions from 49 ASD and 54 control individuals (false discovery rate [FDR] < 0.05 in the "whole cortex" analysis of Gandal et al.[58]). We found that the DEGs in both lines of iPSC models significantly overlapped with the DEGs in the ASD postmortem brain (Figure S20A, $p = 1.81 \times 10^{-7}$ for iPSCs with the chr7:g.26200781C>T variant, $p = 0.0110$ for iPSCs with the chr22:g.19179942C>T variant, hypergeometric test). Subsequently, we analyzed if the DEGs in our iPSC models significantly overlapped with (1) ASD-associated gene modules reported in this ASD postmortem brain study[58] and (2) the developing DLPFC coexpression network gene modules.[59] This analysis identified multiple modules that showed significant overlaps (Figures S20B and S20C, Benjamini-Hochberg-corrected $p < 0.05$). Interestingly, the ASD-associated geneM2 module was commonly identified as the most significant one in both the chr7:g.26200781C>T and the chr22:g.19179942C>T lines. A GO enrichment analysis showed that the terms enriched in this module are small molecule catabolic process and oxidoreductase activity (Figures S20D and S20E; Table S18). The identification of the ASD-associated geneM2 as the module most enriched in the DEGs of both lines suggests that there might be similarities in transcriptomic profiles between the chr7:g.26200781C>T and the chr22:g.19179942C>T lines. To further test this in an unbiased manner, we analyzed the correlation between the FC of each DEG in these two lines. We observed that there is a significant positive correlation across them (Figure S21, $p = 8.65 \times 10^{-8}$, $R = 0.114$). This may reasonably explain why both of these different promoter DNVs are involved in ASD risk.

Together, these results indicate that single promoter DNVs can influence the expression of multiple genes in the TAD containing the DNV and cause alterations in overall transcriptomic profiles that are enriched for genes associated with ASD risk and neurodevelopment in the cases of promoter DNVs analyzed in this study. We also confirmed highly significant correlations of the patterns of transcriptomic changes (FC of each gene) between cells with homozygous and heterozygous variants (Figure S18, $p < 2.2 \times 10^{-16}$ for both chr7:g.26200781C>T and chr22:g.19179942C>T lines). Among DEGs in cells with heterozygous variants, we again observed an overrepresentation of ASD-associated genes, while the significance was attenuated compared with the homozygous knockin cells (Figure S18).

DISCUSSION

In this study, we conceived the idea that TADs might define the impact of rare promoter variants on disease susceptibility and statistically analyzed large WGS datasets (total N = 19,227) of ASD, a disease for which promoter DNVs have recently been reported to be involved in its risk. By meta-analyzing the WGS data from SSC and SPARK, we observed results supporting this hypothesis (Figure 1). When we repeated the analysis using TAD data from various tissues, enrichment of ASD-gene TAD promoter DNVs in the probands was consistently prominent when the datasets from brain and stem cells were used. In particular, we found the strongest enrichment in the analysis using TADs in iPSC-derived neurons. The iPSC-derived neurons are expected to represent a population of pure, recently differentiated neurons, whereas bulk postmortem brain tissues contain various cell types (e.g., glial and vascular cells) and are derived from elderly individuals, except for the cortical plate and germinal zone samples that showed the second and third most significant enrichments. Therefore, our observations are in line with previous studies demonstrating the important role of early developmental neurons in the pathogenesis of ASD.[60–66]

After observing a significant overrepresentation of ASD-gene TAD promoter DNVs in ASD probands, we analyzed the characteristics of these variants (Figures 2 and S3). Among the findings in these analyses, the enrichment of ASD-gene TAD promoter DNVs at CTCF-bound regions in ASD would be of particular interest, as it can provide mechanistic insights into how mutations confer the risk of ASD. According to multiple recent studies,[23,67,68] it has been reported that CTCF contributes not only to the formation of TAD boundaries but also to the stabilization of promoter-enhancer interactions. Given that, it would be reasonable to assume that some of the ASD-gene TAD promoter DNVs in ASD probands may contribute to disease risk by altering these interactions.

Our comprehensive exploration of TADs enriched for promoter DNVs in ASD, the TADWAS, identified two TADs that achieved significance after Bonferroni correction (Figure 3). Also, the contribution of genetic variants in TADs with enrichment of promoter DNVs was supported by the overrepresentation of ASD heritability in promoters of these regions. Further analysis will facilitate our understanding of the functional effects of variants in these regions, which can be either common or rare, underlying their association with ASD.

In the last part of this study, we experimentally evaluated the consequences of two selected promoter DNVs in ASD using human iPSC models (Figure 4). Overall, our results indicate that a single promoter SNV can induce both local gene expression changes within a TAD and global transcriptomic alteration. These differential expressions of non-nearest-neighbor genes align well with the results reported by Kim et al.[69] In their study using iPSC-based models, it was shown that the non-coding DNVs identified in ASD induce expression changes of genes computationally predicted to be targets of chromatin interactions. Notably, many of these genes that showed differential expression were not those nearest to the non-coding DNV. Therefore, their results and ours together highlight the role of non-coding DNVs with gene-regulatory impacts in the genetic architecture of ASD. Also, our observation in this study that manipulation of a single promoter leads to differential expression of multiple genes would have implications for therapeutic strategies for genetic diseases caused by CNVs, such as those in the 22q11.2 locus analyzed in this study. While research and development of gene therapy for monogenic Mendelian diseases supplementing the causal gene is currently being actively conducted,[70–72] a direct adaptation of such interventions to the treatment of CNV syndromes, where usually multiple genes are

disrupted, is not easy. To solve this problem, one approach is to identify the key gene(s) in the CNV region and complement it. In addition to this strategy, according to the findings of this study, modulation of a single promoter to normalize the expression of multiple genes in the disease-responsible region could be a reasonable alternative, while various experimental verifications are mandatory to prove this concept.

In conclusion, our findings show that TADs and their gene contents define the impact of promoter DNVs on ASD risk based on statistical evidence from an analysis of large WGS datasets. In addition, this study suggests that the integration of information on TADs with appropriate gene sets contributes to a better understanding of the role of rare non-coding variants in the genetic architectures of other diseases and traits. This work was enabled by the utilization of large WGS data publicly available under the authorization; various genome annotations, including TAD; and recently established cell and genome engineering techniques. By further scaling up studies in terms of both the sample size and experimental comprehensiveness, we will better understand the genetic architecture and biology of ASD.

Limitations of the study

Considering the limitations of this study and future perspectives, first, although this study meta-analyzes large-scale WGS datasets for ASD currently available, the obtained statistical power might still be insufficient in the context of a category-wide association study.[73] Indeed, the levels of significance observed for ASD-gene TAD promoter DNVs in this study do not reach a stringent significance threshold, given that there are thousands of functional categories of non-coding variants based on combinations of various annotations. Also, the results of association analysis can be influenced by selecting different sets of known disease genes. While in this study we used the SFARI genes representing a general and widely accepted list of ASD-associated genes, to address this, we also performed analyses using a longer list covering neurodevelopmental disorders genes not included in the SFARI genes (see STAR Methods, total n = 1,666 genes) and a shorter list of high-confidence ASD genes only, including the SFARI genes in categories S, 1, and 2 (n = 498 genes). We found that ASD-gene TAD promoter DNVs are consistently more frequent in ASD probands than in unaffected siblings, whereas non-ASD-gene TAD promoter DNVs showed non-significant association, with ORs around 1 (Figure S22; Table S19). Thus, the concept that DNVs in ASD-gene TADs contribute more to the disease risk itself can be considered to have a certain validity. Meanwhile, the limitation in the sample size and the corresponding statistical power also apply to the observations for ASD-gene-linked promoter DNVs. Although we did not find any significant association of this type of promoter DNV with ASD, this might be due to the small number of ASD-gene-linked promoter DNVs identified in our study (Figures 1B–1D). Given that they have the potential to alter the expression of ASD genes, and that the ORs for them were similar to those for ASD-gene TAD promoter DNVs, as shown in Figures 1B–1D, it might be reasonable to assume that ASD-gene-linked promoter DNVs also confer the risk of ASD. Therefore, in any case, it is important to further expand the sample size to confirm the results of this study and to investigate the roles of various types of non-coding rare variants. Second, due to the difficulty in obtaining original DNA samples, the promoter DNVs identified in the SPARK dataset were not confirmed by Sanger or target amplicon sequencing. Nevertheless, this study exactly followed the methods for DNV calling in a previous study of SSC,[16] in which selected promoter DNVs were confirmed by Sanger sequencing with a validation rate of 100% (48/48 DNV calls), and we observed an expected pattern of overrepresentation of LoF DNVs in ASD in SPARK. Based on these, it is generally considered that the accuracy of DNV calls in the SPARK dataset generated in this study is sufficiently high. Third, while we meta-analyzed the results from the SSC and SPARK cohorts to increase statistical power, it has to be taken into account that these two cohorts have different characteristics. For example, the SSC participants were recruited primarily in university-affiliated research clinics, whereas the SPARK cohort includes online-recruited voluntary participants whose phenotypic information and ASD diagnoses were self- or parent-reported.[18] Indeed, the effect sizes for conserved promoter DNVs were quite different between the SSC and the SPARK cohorts, suggesting that there might be a bias due to cohort characteristics. Fourth, the number of promoter DNVs whose effects were experimentally assessed in this study was very small. A more extensive analysis of promoter DNVs is required to obtain the full picture of their functional consequences. Also, knowledge of the impact of variants at more differentiated stages would provide additional insights. The development and improvement of high-throughput screening for functional effects of non-coding variants is a field under active research.[74,75] An application of new technologies will facilitate the understanding of the genetic architectures of complex traits accounted for by promoter DNVs and other types of rare non-coding variants.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human induced pluripotent stem cells
- METHOD DETAILS
  - Studied cohorts
  - DNV calling and filtering in SPARK
  - TADs in human tissues and cell lines
  - Annotation of DNVs
  - Case-control analysis of promoter DNVs
  - Analysis of chromatin states
  - TF binding enrichment analysis
  - Prediction scores for non-coding variants
  - Annotation of TADs
  - Calculation of *de novo* promoter SNV rates for each TAD
  - TADWAS

- LDSC
- Transfection of CRISPR/Cas9 components
- Single-cell sorting
- Screening of knock-in iPSC clones
- Screening of possible off-target variants
- RNA-seq
- Read mapping and generation of count matrices
- DEG and downstream analysis

- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100488.

ACKNOWLEDGMENTS

We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, and E. Wijsman). We are grateful to all of the families in SPARK, the SPARK clinical sites, and SPARK staff. We appreciate obtaining access to phenotypic and genetic data on Simons Foundation Autism Research Initiative (SFARI) Base. Approved researchers can obtain the SSC (https://www.sfari.org/resource/simons-simplex-collection/) and the SPARK (https://www.sfari.org/resource/spark/) population datasets described in this study by applying at https://base.sfari.org. We thank the Support Unit for Bio-Material Analysis, RIKEN CBS Research Resources Division, for technical help. We thank Daisuke Yamaguchi and Hao Zhang from BITS Co., Ltd., for their assistance in the construction of the analytical environment. This research was supported by the RIKEN Center for Brain Science and the following funding agencies: AMED under grant nos. JP20dm0107133 (A.T.), JP22km0405214 (A.T.), and JP20dm0307028 (A.T.) and JSPS KAKENHI under grant nos. JP20H05777 (A.T.), JP21H02855 (A.T.), JP21K15752 (T.N.), and JP22K20751 (S.M.).

AUTHOR CONTRIBUTIONS

A.T. conceived the study. T.N., J.U., S.M., A.K., H.Y., T.H., and A.T. performed data preparation and analyses. T.N. and K.H. performed iPSC experiments. J.U. played a leading role in the bioinformatics and statistical analyses. A.T. wrote the first draft, and T.N., J.U., S.M., A.K., and A.T. jointly revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 14, 2023
Revised: August 24, 2023
Accepted: January 2, 2024
Published: January 26, 2024

REFERENCES

1. Maenner, M.J., Shaw, K.A., Bakian, A.V., Bilder, D.A., Durkin, M.S., Esler, A., Furnier, S.M., Hallas, L., Hall-Lande, J., Hudson, A., et al. (2021). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. MMWR. Surveill. Summ. *70*, 1–16.
2. Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat. Genet. *46*, 881–885.
3. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444.
4. Matoba, N., Liang, D., Sun, H., Aygün, N., McAfee, J.C., Davis, J.E., Raffield, L.M., Qian, H., Piven, J., Li, Y., et al. (2020). Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. Transl. Psychiatry *10*, 265.
5. Antaki, D., Guevara, J., Maihofer, A.X., Klein, M., Gujral, M., Grove, J., Carey, C.E., Hong, O., Arranz, M.J., Hervas, A., et al. (2022). A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat. Genet. *54*, 1284–1292.
6. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
7. C Yuen, R.K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R.V., Whitney, J., Deflaux, N., Bingham, J., Wang, Z., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat. Neurosci. *20*, 602–611.
8. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568–584.e23.
9. Takata, A., Miyake, N., Tsurusaki, Y., Fukai, R., Miyatake, S., Koshimizu, E., Kushima, I., Okada, T., Morikawa, M., Uno, Y., et al. (2018). Integrative Analyses of De Novo Mutations Provide Deeper Biological Insights into Autism Spectrum Disorder. Cell Rep. *22*, 734–747.
10. Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J.B., Obiajulu, J.U., Wright, J.R., Murali, S.C., Xu, S.X., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat. Genet. *54*, 1305–1319.
11. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331.
12. Du, X., Gao, X., Liu, X., Shen, L., Wang, K., Fan, Y., Sun, Y., Luo, X., Liu, H., Wang, L., et al. (2018). Genetic Diagnostic Evaluation of Trio-Based Whole Exome Sequencing Among Children With Diagnosed or Suspected Autism Spectrum Disorder. Front. Genet. *9*, 594.
13. Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T.N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L.G., Muzny, D.M., et al. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. NPJ Genom. Med. *4*, 19.
14. Tammimies, K., Marshall, C.R., Walker, S., Kaur, G., Thiruvahindrapuram, B., Lionel, A.C., Yuen, R.K.C., Uddin, M., Roberts, W., Weksberg, R., et al. (2015). Molecular Diagnostic Yield of Chromosomal Microarray Analysis and Whole-Exome Sequencing in Children With Autism Spectrum Disorder. JAMA *314*, 895–903.
15. Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. JAMA *318*, 1182–1184.
16. An, J.Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576.
17. Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A., et al. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. Cell *171*, 710–722.e12.

18. SPARK Consortium Electronic address pfeliciano@simonsfoundationorg; SPARK Consortium (2018). SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. Neuron *97*, 488–493.

19. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

20. Kong, A., Frigge, M.L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S.A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., et al. (2012). Rate of de novo mutations and the importance of father's age to disease risk. Nature *488*, 471–475.

21. Goldmann, J.M., Wong, W.S.W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E.L.M., Hoischen, A., Roach, J.C., et al. (2016). Parent-of-origin-specific signatures of de novo mutations. Nat. Genet. *48*, 935–939.

22. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism. *4*, 36.

23. Oh, S., Shao, J., Mitra, J., Xiong, F., D'Antonio, M., Wang, R., Garcia-Bassets, I., Ma, Q., Zhu, X., Lee, J.H., et al. (2021). Enhancer release and retargeting activates disease-susceptibility genes. Nature *595*, 735–740.

24. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*, eaat8464.

25. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

26. Roadmap Epigenomics Consortium; Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

27. Zou, Z., Ohta, T., Miura, F., and Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. Nucleic Acids Res. *50*, W175–W182.

28. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

29. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

30. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. *48*, 214–220.

31. Huang, Y.F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet. *49*, 618–624.

32. Zhang, S., He, Y., Liu, H., Zhai, H., Huang, D., Yi, X., Dong, X., Wang, Z., Zhao, K., Zhou, Y., et al. (2019). regBase: whole genome base-wise aggregation and functional prediction for human non-coding regulatory variants. Nucleic Acids Res. *47*, e134.

33. di Iulio, J., Bartha, I., Wong, E.H.M., Yu, H.C., Lavrenko, V., Yang, D., Jung, I., Hicks, M.A., Shah, N., Kirkness, E.F., et al. (2018). The human noncoding genome defined by genetic diversity. Nat. Genet. *50*, 333–337.

34. Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.H., and Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci. Rep. *5*, 10576.

35. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.

36. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151.

37. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

38. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

39. Sullivan, P.F., Agrawal, A., Bulik, C.M., Andreassen, O.A., Børglum, A.D., Breen, G., Cichon, S., Edenberg, H.J., Faraone, S.V., Gelernter, J., et al. (2018). Psychiatric Genomics: An Update and an Agenda. Am. J. Psychiatry *175*, 15–27.

40. Demontis, D., Walters, R.K., Martin, J., Mattheisen, M., Als, T.D., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M., et al. (2019). Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat. Genet. *51*, 63–75.

41. Walters, R.K., Polimanti, R., Johnson, E.C., McClintick, J.N., Adams, M.J., Adkins, A.E., Aliev, F., Bacanu, S.A., Batzler, A., Bertelsen, S., et al. (2018). Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat. Neurosci. *21*, 1656–1669.

42. Wightman, D.P., Jansen, I.E., Savage, J.E., Shadrin, A.A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B.S., Drange, O.K., et al. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. Nat. Genet. *53*, 1276–1282.

43. Watson, H.J., Yilmaz, Z., Thornton, L.M., Hübel, C., Coleman, J.R.I., Gaspar, H.A., Bryois, J., Hinney, A., Leppä, V.M., Mattheisen, M., et al. (2019). Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat. Genet. *51*, 1207–1214.

44. Mullins, N., Forstner, A.J., O'Connell, K.S., Coombes, B., Coleman, J.R.I., Qiao, Z., Als, T.D., Bigdeli, T.B., Børte, S., Bryois, J., et al. (2021). Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat. Genet. *53*, 817–829.

45. Johnson, E.C., Demontis, D., Thorgeirsson, T.E., Walters, R.K., Polimanti, R., Hatoum, A.S., Sanchez-Roige, S., Paul, S.E., Wendt, F.R., Clarke, T.K., et al. (2020). A large-scale genome-wide association study meta-analysis of cannabis use disorder. Lancet Psychiatr. *7*, 1032–1045.

46. Howard, D.M., Adams, M.J., Clarke, T.K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R.I., Hagenaars, S.P., Ward, J., Wigmore, E.M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. *22*, 343–352.

47. International Obsessive Compulsive Disorder Foundation Genetics Collaborative IOCDF-GC and OCD Collaborative Genetics Association Studies OCGAS (2018). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. Mol. Psychiatry *23*, 1181–1188.

48. Forstner, A.J., Awasthi, S., Wolf, C., Maron, E., Erhardt, A., Czamara, D., Eriksson, E., Lavebratt, C., Allgulander, C., Friedrich, N., et al. (2021).

Genome-wide association study of panic disorder reveals genetic overlap with neuroticism and depression. Mol. Psychiatry *26*, 4179–4190.

49. Nievergelt, C.M., Maihofer, A.X., Klengel, T., Atkinson, E.G., Chen, C.Y., Choi, K.W., Coleman, J.R.I., Dalvie, S., Duncan, L.E., Gelernter, J., et al. (2019). International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. Nat. Commun. *10*, 4558.

50. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508.

51. Yu, D., Sul, J.H., Tsetsos, F., Nawaz, M.S., Huang, A.Y., Zelaya, I., Illmann, C., Osiecki, L., Darrow, S.M., Hirschtritt, M.E., et al. (2019). Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. Am. J. Psychiatry *176*, 217–227.

52. Gazal, S., Loh, P.R., Finucane, H.K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A.L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat. Genet. *50*, 1600–1607.

53. Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature *604*, 310–315.

54. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., and Eichler, E.E. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat. Genet. *51*, 106–116.

55. Chun, S., Westmoreland, J.J., Bayazitov, I.T., Eddins, D., Pani, A.K., Smeyne, R.J., Yu, J., Blundon, J.A., and Zakharenko, S.S. (2014). Specific disruption of thalamic inputs to the auditory cortex in schizophrenia models. Science *344*, 1178–1182.

56. Khan, T.A., Revah, O., Gordon, A., Yoon, S.J., Krawisz, A.K., Goold, C., Sun, Y., Kim, C.H., Tian, Y., Li, M.Y., et al. (2020). Neuronal defects in a human cellular model of 22q11.2 deletion syndrome. Nat. Med. *26*, 1888–1898.

57. Stark, K.L., Xu, B., Bagchi, A., Lai, W.S., Liu, H., Hsu, R., Wan, X., Pavlidis, P., Mills, A.A., Karayiorgou, M., and Gogos, J.A. (2008). Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. Nat. Genet. *40*, 751–760.

58. Gandal, M.J., Haney, J.R., Wamsley, B., Yap, C.X., Parhami, S., Emani, P.S., Chang, N., Chen, G.T., Hoftman, G.D., de Alba, D., et al. (2022). Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. Nature *611*, 532–539.

59. Werling, D.M., Pochareddy, S., Choi, J., An, J.Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A.M.M., et al. (2020). Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. Cell Rep. *31*, 107489.

60. Marchetto, M.C., Belinson, H., Tian, Y., Freitas, B.C., Fu, C., Vadodaria, K., Beltrao-Braga, P., Trujillo, C.A., Mendes, A.P.D., Padmanabhan, K., et al. (2017). Altered proliferation and networks in neural cells derived from idiopathic autistic individuals. Mol. Psychiatry *22*, 820–835.

61. Sugathan, A., Biagioli, M., Golzio, C., Erdin, S., Blumenthal, I., Manavalan, P., Ragavendran, A., Brand, H., Lucente, D., Miles, J., et al. (2014). CHD8 regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. Proc. Natl. Acad. Sci. USA *111*, E4468–E4477.

62. Cope, E.C., Briones, B.A., Brockett, A.T., Martinez, S., Vigneron, P.A., Opendak, M., Wang, S.S., and Gould, E. (2016). Immature Neurons and Radial Glia, But Not Astrocytes or Microglia, Are Altered in Adult Cntnap2 and Shank3 Mice, Models of Autism. eNeuro *3*.

63. Avino, T.A., Barger, N., Vargas, M.V., Carlson, E.L., Amaral, D.G., Bauman, M.D., and Schumann, C.M. (2018). Neuron numbers increase in the human amygdala from birth to adulthood, but not in autism. Proc. Natl. Acad. Sci. USA *115*, 3710–3715.

64. Kang, S.C., Jaini, R., Hitomi, M., Lee, H., Sarn, N., Thacker, S., and Eng, C. (2020). Decreased nuclear Pten in neural stem cells contributes to deficits in neuronal maturation. Mol. Autism. *11*, 43.

65. Paulsen, B., Velasco, S., Kedaigle, A.J., Pigoni, M., Quadrato, G., Deo, A.J., Adiconis, X., Uzquiano, A., Sartore, R., Yang, S.M., et al. (2022). Autism genes converge on asynchronous development of shared neuron classes. Nature *602*, 268–273.

66. Schafer, S.T., Paquola, A.C.M., Stern, S., Gosselin, D., Ku, M., Pena, M., Kuret, T.J.M., Liyanage, M., Mansour, A.A., Jaeger, B.N., et al. (2019). Pathological priming causes developmental gene network heterochronicity in autistic subject-derived neurons. Nat. Neurosci. *22*, 243–255.

67. Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D.R., and Zhao, K. (2017). CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. Mol. Cell *67*, 1049–1058.e6.

68. Kubo, N., Ishii, H., Xiong, X., Bianco, S., Meitinger, F., Hu, R., Hocker, J.D., Conte, M., Gorkin, D., Yu, M., et al. (2021). Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat. Struct. Mol. Biol. *28*, 152–161.

69. Kim, I.B., Lee, T., Lee, J., Kim, J., Lee, S., Koh, I.G., Kim, J.H., An, J.Y., Lee, H., Kim, W.K., et al. (2022). Non-coding de novo mutations in chromatin interactions are implicated in autism spectrum disorder. Mol. Psychiatry *27*, 4680–4694.

70. Dunbar, C.E., High, K.A., Joung, J.K., Kohn, D.B., Ozawa, K., and Sadelain, M. (2018). Gene therapy comes of age. Science *359*, eaan4672.

71. Sztainberg, Y., and Zoghbi, H.Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. Nat. Neurosci. *19*, 1408–1417.

72. Kim, J., Hu, C., Moufawad El Achkar, C., Black, L.E., Douville, J., Larson, A., Pendergast, M.K., Goldkind, S.F., Lee, E.A., Kuniholm, A., et al. (2019). Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. N. Engl. J. Med. *381*, 1644–1652.

73. Werling, D.M., Brand, H., An, J.Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet. *50*, 727–736.

74. Townsley, K.G., Brennand, K.J., and Huckins, L.M. (2020). Massively parallel techniques for cataloguing the regulome of the human brain. Nat. Neurosci. *23*, 1509–1521.

75. Kinney, J.B., and McCandlish, D.M. (2019). Massively Parallel Assays and Quantitative Sequence-Function Relationships. Annu. Rev. Genomics Hum. Genet. *20*, 99–127.

76. Highnam, G., Wang, J.J., Kusler, D., Zook, J., Vijayan, V., Leibovich, N., and Mittelman, D. (2015). An analytical framework for optimizing variant discovery from personal genomes. Nat. Commun. *6*, 6275.

77. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

78. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.

79. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

80. Wei, Q., Zhan, X., Zhong, X., Liu, Y., Han, Y., Chen, W., and Li, B. (2015). A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics *31*, 1375–1381.

81. Cingolani, P., Patel, V.M., Coon, M., Nguyen, T., Land, S.J., Ruden, D.M., and Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. Front. Genet. *3*, 35.

82. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

83. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

84. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

85. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

86. Balduzzi, S., Rücker, G., and Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. Evid. Based. Ment. Health *22*, 153–160.

87. Cinar, O., and Viechtbauer, W. (2022). The poolr Package for Combining Independent and Dependent p Values. J. Stat. Softw. *101*, 1–42.

88. Rogers, M.F., Shihab, H.A., Gaunt, T.R., and Campbell, C. (2017). CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. Sci. Rep. *7*, 11597.

89. Rogers, M.F., Gaunt, T.R., and Campbell, C. (2020). CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. Bioinformatics *36*, 3637–3644.

90. Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics *31*, 761–763.

91. Yang, H., Chen, R., Wang, Q., Wei, Q., Ji, Y., Zheng, G., Zhong, X., Cox, N.J., and Li, B. (2019). De novo pattern discovery enables robust assessment of functional consequences of non-coding variants. Bioinformatics *35*, 1453–1460.

92. Shihab, H.A., Rogers, M.F., Gough, J., Mort, M., Cooper, D.N., Day, I.N.M., Gaunt, T.R., and Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics *31*, 1536–1543.

93. Rogers, M.F., Shihab, H.A., Mort, M., Cooper, D.N., Gaunt, T.R., and Campbell, C. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics *34*, 511–513.

94. Ioannidis, N.M., Davis, J.R., DeGorter, M.K., Larson, N.B., McDonnell, S.K., French, A.J., Battle, A.J., Hastie, T.J., Thibodeau, S.N., Montgomery, S.B., et al. (2017). FIRE: functional inference of genetic variants that regulate gene expression. Bioinformatics *33*, 3895–3901.

95. Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

96. Gulko, B., and Siepel, A. (2018). How Much Information is Provided by Human Epigenomic Data? An Evolutionary View. Preprint at bioRxiv. https://doi.org/10.1101/317719.

97. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X.J., Yip, K.Y., Khurana, E., and Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol. *15*, 480.

98. Wells, A., Heckerman, D., Torkamani, A., Yin, L., Sebat, J., Ren, B., Telenti, A., and di Iulio, J. (2019). Ranking of non-coding pathogenic variants and putative essential regions of the human genome. Nat. Commun. *10*, 5241.

99. Gussow, A.B., Copeland, B.R., Dhindsa, R.S., Wang, Q., Petrovski, S., Majoros, W.H., Allen, A.S., and Goldstein, D.B. (2017). Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. PLoS One *12*, e0181604.

100. Zhou, L., and Zhao, F. (2018). Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med. *10*, 53.

101. Smedley, D., Schubach, M., Jacobsen, J.O.B., Köhler, S., Zemojtel, T., Spielmann, M., Jäger, M., Hochheiser, H., Washington, N.L., McMurry, J.A., et al. (2016). A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am. J. Hum. Genet. *99*, 595–606.

102. D Turner, S. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J. Open Source Softw. *3*, 731.

103. Krueger, F., James, F., Ewels, P., Afyounian, E., and Schuster-Boeckler, B. (2021). FelixKrueger/TrimGalore: v0.6.7. via Zenodo.

104. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

105. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

106. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

107. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

108. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

109. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523.

110. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

111. Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H., and Valen, E. (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. *47*, W171–W174.

112. Bae, S., Park, J., and Kim, J.S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics *30*, 1473–1475.

113. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

114. Won, H., de la Torre-Ubieta, L., Stein, J.L., Parikshak, N.N., Huang, J., Opland, C.K., Gandal, M.J., Sutton, G.J., Hormozdiari, F., Lu, D., et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. Nature *538*, 523–527.

115. Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43*, W589–W598.

116. Kaplanis, J., Samocha, K.E., Wiel, L., Zhang, Z., Arvai, K.J., Eberhardt, R.Y., Gallone, G., Lelieveld, S.H., Martin, H.C., McRae, J.F., et al.

(2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature *586*, 757–762.

117. Nyholt, D.R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am. J. Hum. Genet. *74*, 765–769.

118. Short, P.J., McRae, J.F., Gallone, G., Sifrim, A., Won, H., Geschwind, D.H., Wright, C.F., Firth, H.V., FitzPatrick, D.R., Barrett, J.C., and Hurles, M.E. (2018). De novo mutations in regulatory elements in neurodevelopmental disorders. Nature *555*, 611–616.

119. Brandão, K.O., Grandela, C., Yiangou, L., Mummery, C.L., and Davis, R.P. (2022). CRISPR/Cas9-Mediated Introduction of Specific Heterozygous Mutations in Human Induced Pluripotent Stem Cells. Methods Mol. Biol. *2454*, 531–557.

120. Love, M.I., Anders, S., Kim, V., and Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Res. *4*, 1070.

121. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
iMatrix-511 silk	Takara Bio Inc.	Cat# 892011
TrypLE™ SELECT	Thermo Fisher Scientific	Cat# 12563029
0.5 mM EDTA/PBS solution	NACALAI TESQUE, INC.	Cat# 13567-84
StemFit AK02N	RIPROCELL	Cat# RCAK02N
CultureSure Y-27632	Fujifilm	Cat# 036-24023
Alt-R S.p. HiFi Cas9 Nuclease V3	Integrated DNA Technologies	Cat# 1081060
7-AAD	BD Biosciences	Cat# 559925
Primocin(TM)	NACALAI TESQUE	Cat# 14860-36
Accutase	NACALAI TESQUE	Cat# 12679-54
BigDye Terminator V3.1	Thermo Fisher Scientific	Cat# 4337455
STEM CELL BANKER GMP grade	Takara Bio Inc.	Cat# 11924
Trizol reagent	Thermo Fisher Scientific	Cat# 15596026
Recombinant DNaseI (Rnase-free)	Takara Bio Inc.	Cat# 2270A
Critical commercial assays		
Guide-it™ sgRNA *In vitro* Transcription Kit	Takara Bio Inc.	Cat# 632635
NEBNext Ultra RNA Library Prep Kit for Illumina	New England BioLabs Inc.	Cat # E7530
Deposited data		
RNA-seq data of wild-type/mutant iPSCs	This study	The NDBC Human Database, Japan (Accession number: JGA: JGAS000651)
Population datasets	SFARI Base	https://www.sfari.org/resource/sfari-base/
TAD list of DLPFC	PsychENCODE Integrative Analysis resource	http://resource.psychencode.org/Datasets/Derived/DER-18_TAD_adultbrain.bed
TAD list of iPSC-derived neurons	GEO database	GSE79965, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79965
TAD lists of germinal zone and cortical plate	GEO database	GSE77565, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77565
Other TAD lists	3D Genome Browser[36]	http://3dgenome.fsm.northwestern.edu/
Data of reference epigenomics	Roadmap Epigenomics Project[26]	http://www.roadmapepigenomics.org/
Data of enhancer regions	PsychENCODE[24]	http://resource.psychencode.org/Datasets/Derived/DER-18_TAD_adultbrain.bed
GWAS sumstat data	Psychiatric Genomics Consortium; Matoba et al.[4]; Alkes Price's lab	https://www.med.unc.edu/pgc/download-results/; https://bitbucket.org/steinlabunc/spark_asd_sumstats/src/master/ASD_SPARK_iPSYCH_PGC.tsv.gz/; https://alkesgroup.broadinstitute.org/LDSCORE/independent_sumstats/
Experimental models: Organisms/strains		
Human: 201B7-Ff	RIKEN BioResource Research Center	HPS4290
Oligonucleotides		
Oligo lists	This paper Tables S9, S10, and S11	N/A
Software and algorithms		
R	The R Foundation	https://www.r-project.org/
BWA-MEM (v.0.7.15)	Highnam et al.[76]	https://github.com/kaist-ina/BWA-MEME/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GATK (v.3.5 and v4.1.9.0)	DePristo et al.[77]	https://gatk.broadinstitute.org/hc/
bcftools (v.1.10.2)	Danecek et al.[78]	http://samtools.github.io/bcftools/
vcftools (v.0.1.17)	Danecek et al.[79]	https://vcftools.sourceforge.net/
TrioDenovo (v.0.06)	Wei et al.[80]	https://genome.sph.umich.edu/wiki/Triodenovo
SnpSift (v.5.0c)	Cingolani et al.[81]	http://pcingola.github.io/SnpEff/
RepeatMasker	UCSC	https://www.repeatmasker.org/
PLINK (v.1.90b6.24)	Purcell et al.[82]; Chang et al.[83]	https://www.cog-genomics.org/plink/
3D Genome Browser	Wang et al.[36]	http://3dgenome.fsm.northwestern.edu/
Ensembl Variant Effect Predictor	McLaren et al.[84]	https://asia.ensembl.org/info/docs/tools/vep/index.html
bedtools	Quinlan et al.[85]	https://bedtools.readthedocs.io/
meta package (v5.5-0)	Balduzzi et al.[86]	https://cran.r-project.org/
poolr package (v.1.1.1)	Cinar et al.[87]	https://cran.r-project.org/
regBase (v.1.1.1)	Zhang et al.[32]	https://github.com/mulinlab/regBase/
CADD (v.1.3, 1.4, and 1.6)	Kircher et al.[28]; Rentzsch et al.[29]	https://cadd.gs.washington.edu/
CDTS	di Iulio et al.[33]	http://www.hli-opendata.com/noncoding/
Cscape	Rogers et al.[88]	http://CScape.biocompute.org.uk/
Cscape_Somatic	Rogers et al.[89]	http://cscape-somatic.biocompute.org.uk/
DANN	Quang et al.[90]	https://cbcl.ics.uci.edu/public_data/DANN/
DVAR	Yang et al.[91]	https://www.vumc.org/cgg/dvar
Eigen/Eigen_PC	Ionita-Laza et al.[30]	http://www.columbia.edu/~ii2135/eigen.html
FATHMM (MKL and XF)	MKL, Shihab et al.[92]; XF,Rogers et al.[93]	http://fathmm.biocompute.org.uk/, http://fathmm.biocompute.org.uk/fathmm-xf/
FIRE	Ioannidis et al.[94]	https://sites.google.com/site/fireregulatoryvariation/
fitCons	Gulko et al.[95]	http://compgen.cshl.edu/fitCons/
FitCons2	Gulko et al.[96]	https://github.com/CshlSiepelLab/FitCons2
FunSeq2	Fu et al.[97]	http://funseq2.gersteinlab.org/
GenoCanyon	Lu et al.[34]	https://zhaocenter.org/GenoCanyon_Index.html
LINSIGHT	Huang et al.[31]	https://github.com/CshlSiepelLab/LINSIGHT
ncER	Wells et al.[98]	https://github.com/TelentiLab/ncER_datasets
Orion	Gussow et al.[99]	https://github.com/igm-team/orion-public
PAFA	Zhou et al.[100]	http://159.226.67.237:8080/pafa
ReMM	Smedley et al.[101]	https://remm.bihealth.org/
qqman package	Tuner et al.[102]	https://cran.r-project.org/
LDSC (v.1.0.1)	Bulik-Sullivan et al.[37]; Finucane et al.[38]	https://github.com/bulik/ldsc
trim_galore (v.0.6.6)	Krueger et al.[103]	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
STAR (v.2.7.9a)	Dobin et al.[104]	https://github.com/alexdobin/STAR/
Samtools (v.1.3.1)	Li et al.[105]	http://www.htslib.org/
featureCounts (v.2.0.1)	Liao et al.[106]	https://subread.sourceforge.net/
DESeq2 package (v.1.36.0)	Love et al.[107]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
sva package (v.3.44.0)	Leek et al.[108]	https://bioconductor.org/packages/release/bioc/html/sva.html
Metascape	Zhou et al.[109]	https://metascape.org/
Cytoscape (v.3.7.2)	Shannon et al.[110]	https://cytoscape.org/
CHOPCHOP	Labun et al.[111]	https://chopchop.cbu.uib.no/
Cas-OFFinder	Bae et al.[112]	http://www.rgenome.net/cas-offinder/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GSEA (v.4.3.2)	Subramanian et al.[113]	https://www.gsea-msigdb.org/gsea/
Analysis codes used in this manuscript	This paper	https://doi.org/10.5281/zenodo.10437260 or upon request
Other		
Neon™ Transfection System	Thermo Fisher Scientific	Cat# MPK1025
BD FACSAria II	BD Biosciences	N/A
ABI 3730xl sequencer	Life Technologies	Cat# 3730xlLV-100
Agilent 2100 Bioanalyzer	Agilent Technologies, Ltd.	N/A
Illumina NovaSeq 6000 System	Illumina, Inc.	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Atsushi Takata (atsushi.takata@riken.jp).

Materials availability

All unique/stable reagents generated in this study are available from the lead contact with a completed Materials Transfer Agreement.

Data and code availability

- The population datasets in this study are available from SFARI Base (https://www.sfari.org/resource/sfari-base/). The interval lists of TADs in various tissues and cell lines are available through the following resources: human DLPFC in the PsychENCODE Integrative Analysis resource (http://resource.psychencode.org/Datasets/Derived/DER-18_TAD_adultbrain.bed), iPSC-derived neurons in the GEO database (GEO accession number: GSE79965), germinal zone and cortical plate in GEO database (GEO accession number: GSE77565) of ref[114], respectively, and the others from the 3D Genome Browser[36] (http://3dgenome.fsm.northwestern.edu/). The data of reference epigenomes across 127 tissues/cells are available from the Roadmap Epigenomics Project[26] (http://www.roadmapepigenomics.org/). The data of enhancer regions in this study is available from PsychENCODE[24] (DER-04a_hg38lft_PEC_enhancers). GWAS summary statistics for various psychiatric disorders, including ASD, and traits analyzed in the UK Biobank projects are available from the download site of Psychiatric Genomics Consortium (PGC, https://www.med.unc.edu/pgc/download-results/), an ASD GWAS meta-analysis study (https://bitbucket.org/steinlabunc/spark_asd_sumstats/src/master/ASD_SPARK_iPSYCH_PGC.tsv.gz/), and the download site of Alkes Price's Lab (https://alkesgroup.broadinstitute.org/LDSCORE/independent_sumstats/).
- Original codes used in this study are available at https://doi.org/10.5281/zenodo.10437260 or upon request.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- RNA-seq data of wild-type and mutant iPSCs generated in this study are available at the National Bioscience Database Center (NDBC) Human Database, Japan (Accession number: JGA: JGAS000651).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human induced pluripotent stem cells

The human iPSCs (201B7-Ff [HPS4290], supplied by RIKEN BioResource Research Center) were maintained in feeder-free conditions. The HPS4290 iPSCs were established from the skin sample of 36 years old Caucasian female, and the clinical information was not available (https://cellbank.brc.riken.jp/cell_bank/CellInfo/?cellNo=HPS4290&lang=En). The procedure for the passage is as follows. One day before the passage, 0.5 μg/cm^2 of iMatrix-511 silk (Takara Bio Inc., Shiga, Japan) was coated onto a 10 cm polystyrene dish (Corning Incorporated, NY, USA), and the coated dish was kept at 4°C. The confluent iPSC colonies were separated by incubation with 0.5× TrypLE SELECT (Thermo Fisher Scientific, MA, USA) diluted by 0.5 mM EDTA/PBS solution (NACALAI TESQUE, INC., Kyoto, Japan) at 37°C for 7 min. The 80,000 of the separated single iPSCs were put onto the 10 cm dish coated by iMatrix-511 silk in the condition of StemFit AK02N medium (RIPROCELL, Kanagawa, Japan) including 10 μM of CultureSure Y-27632 (Fujifilm, Tokyo, Japan). On days 1, 4, 5, and 6 after the passage, the medium change was conducted without Y-27632. Cell authentication have not been performed in this study. We confirm that this study complies with all relevant ethical regulations and is reviewed and approved by the ethics committee of RIKEN (approval numbers: W2022-032 and Wako1 2021-001(4)).

METHOD DETAILS

Studied cohorts

We analyzed the data of promoter DNVs in the Simons Simplex Collection (SSC; 1,902 ASD probands, 1,902 unaffected siblings, and their parents) and the Simons Foundation Powering Autism Research for Knowledge (SPARK; 3,153 ASD probands, 2,197 unaffected siblings, and their parents, before the filtering described below) datasets. The data of DNVs in SSC were obtained from a previous publication by An et al.[16] (Table S2 of the corresponding publication). There is a total of 6,787 promoter DNMs (GENCODE_Promoter = 1 in the Table) in SSC. The list of promoter DNVs in SPARK was generated from the cohort VCF files in the WGS1/2/3/4 datasets using the methods described below. All participants provided informed consent through the SSC and SPARK projects.

DNV calling and filtering in SPARK

Candidate DNVs in the SPARK dataset were extracted from the cohort VCF files (WGS1/2/3/4 datasets) obtained from SFARI Base upon authorization. Briefly, the cohort VCF files available through SFARI Base were generated through library preparation with the Illumina PCR-free library protocol, sequencing by Illumina HiSeq X (WGS1/2/3) or NovaSeq 6000 (WGS4), read mapping by BWA-MEM[76] (v0.7.15), individual-level genotyping by GATK[77] (v3.5) HaplotypeCaller with the gVCF mode, and joint genotyping by GATK CombineGVCFs and GenotypeGVCFs. We first normalized the multiallelic variant calls to biallelic and extracted calls with PASS flag using bcftools[78] (v1.10.2). Variant calls whose alternative allele was marked with "*" during the normalization were excluded. We then removed the variant calls at low complexity regions defined by the mdust program (https://github.com/lh3/mdust) using vcftools[79] (v0.1.17). DNV candidates were extracted by GATK (v4.1.9.0) SelectVariants with the –mendelian-violation argument and screened by TrioDenovo[80] (v0.06) with default parameters. Then the selected candidates were hard-filtered using SnpSift[81] (v5.0c) with the same parameters used in the An et al. study[16] to ensure consistency across datasets as follows: for SNVs, FILTER = 'PASS', GQ >= 99, QUAL >= 200, SOR <= 2.5, DP >= 10, ReadPosRankSum >= -1.4, QD >= 3, AB >= 0.22, MQ >= 60, parents' GQ >= 30, parents' AB < 0.05 and Gene Quality Mean >= 50. For Indels, FILTER = 'PASS', GQ >= 99, QUAL >= 200, SOR <= 3, DP >= 10, ReadPosRankSum >= -1.7, QD >= 4, parents' GQ >= 30, parents' DP >= 16, MQ >= 50, DP <= 50, Gene Quality Mean >= 40, and AB > 0.20. For Indels, we excluded variant calls in the repetitive regions defined by RepertMasker (downloaded from the UCSC Table Browser) to extract high-confidence DNVs. As individual-level quality controls, 1) for samples redundantly analyzed in multiple batches, the data from an earlier release were removed, 2) family relationships were confirmed by estimating identity by descent (pi-hat) using PLINK[82,83] (v1.90b6.24), and 3) nine ASD probands and three unaffected siblings who are outliers in the total number of DNVs were identified by repeated Smirnov-Grubbs tests (threshold P = 0.05) and excluded. After applying these filtering steps, we found that there are still a certain number of DNV candidates present in the overall cohort with relatively high minor allele frequencies (MAF, e.g. > 1%). We, therefore, calculated the MAF of DNV candidates in the parents and excluded those with MAF in parents >= 0.5% in the downstream analyses, considering that these DNV candidates are likely to be non-pathogenic or sequencing errors. In the final set of DNVs that passed the above filters, the average ± SD of DNV counts in 3,142 ASD probands and 2,193 unaffected siblings were 59.5 ± 13.0 and 59.7 ± 13.1, respectively, which are comparable to the numbers in SSC (Figure S1).

TADs in human tissues and cell lines

The interval lists of TADs (bed format) in human tissues and cells were obtained using the following resources: DLPFC from the PsychENCODE Integrative Analysis resource[24] (http://resource.psychencode.org/Datasets/Derived/DER-18_TAD_adultbrain.bed), iPSC-derived neurons from the Gene Expression Omnibus (GEO) database (GEO accession number: GSE79965), germinal zone and cortical plate from the GEO database (GEO accession number: GSE77565) of ref[114], respectively, and the others from 3D Genome Browser[36] (http://3dgenome.fsm.northwestern.edu/). The positions of TADs were based on the GRCh38/hg38 assembly, except for the DLPFC, iPSC-derived neuron, germinal zone, and cortical plate datasets, which were hg19-based. The TAD intervals in hg19 were liftoverd to GRCh38/hg38 using the UCSC webtool (http://genome.ucsc.edu/cgi-bin/hgLiftOver), disabling the "Allowing multiple output regions" option. A total of 2,636 out of the 2,735 TADs in DLPFC, 3,124 out of 3,132 in iPSC-derived neurons, 2,371 out of 2,465 in germinal zone, and 2,194 out of 2,286 in cortical plate were converted into GRCh38/hg38. The positions of TSS of protein-coding genes were obtained through Ensemble BioMart.[115]

Annotation of DNVs

For SSC data, we used the annotation in the An et al. study[16] (Table S2 of the corresponding paper) made by using the Ensembl Variant Effect Predictor (VEP)[84] release 90 based on the GENCODE v27 gene definition. To analyze the two datasets in an equal condition, we annotated the variants in SPARK using the same versions of VEP and the GENCODE model. In both datasets, promoter variants are defined as those within 2,000bp upstream of the TSS and with an annotation of "upstream_gene_variant" in the VEP outputs. Variants with an annotation of "frameshift_variant", "splice_acceptor_variant", "splice_donor_variant", "stop_gained", "stop_lost" are considered LoF, and those with "synonymous_variant" are considered synonymous. Genomic positions with PhyloP score >= 2 and/or PhastCons score >= 0.2 (based on 46 vertebrate species) are considered evolutionarily conserved. Annotation of TADs was performed with the above-described intervals of human DLPFC TADs in hg38 by the intersect function of bedtools.[85]

Known ASD genes were defined as those registered in the SFARI Gene database[22] (categories S, 1, 2, and 3, n = 1,010, accessed June 2021) in the main analyses. Their information was annotated based on the Ensembl gene IDs. In the analysis using alternative genesets shown in Figure S22 and Table S19, we used a shorter list of high-confidence ASD genes including SFARI genes in categories S, 1, and 2 (n = 498) and a longer list that includes developmental disorder genes in the following datasets: DDG2P (n = 796, mode of inheritance = "monoallelic", "x-linked" or "both [monoallelic and biallelic]", v2.2, downloaded from https://panelapp.genomicsengland.co.uk/panels/484/), the study by Kaplanis et al.[116] (n = 285), and the study by Fu et al.[11] (n = 252). After removing duplicated genes, the longer list includes 1,666 genes.

Case-control analysis of promoter DNVs

Based on the above annotations and the interval lists of TADs obtained from each dataset, promoter DNVs were classified into the following three groups: ASD-gene linked (the gene immediately downstream of a promoter DNV is an ASD gene), ASD gene TAD (one or more ASD genes are included among the genes within the TAD encompassing a promoter DNV other than the immediately downstream gene; the immediately downstream gene is either an ASD gene or non-ASD gene), and non-ASD gene TAD (no ASD gene is included among the genes within the TAD encompassing a promoter DNV other than the immediately downstream gene; the immediately downstream gene is either an ASD gene or non-ASD gene). We analyzed if the number of each class of promoter DNVs is associated with the case-control labels by logistic regression analysis with the ages of the father and mother at the birth of the child as covariates. In this analysis, we excluded one family (two ASD and one unaffected children, family ID: SF0220207) with a father whose ages at children's birth exceed 1,100 months in the SPARK cohort. When the age of the parents at the time of the child's birth was unknown, the median of the father's or mother's age in each cohort was used. In the final dataset for the case-control analysis, 1,902 probands and 1,902 unaffected siblings in SSC and 3,142 probands and 2,193 unaffected siblings in SPARK were included. A meta-analysis of the two datasets was performed by inverse variance method with a fixed effect model using the meta package (v5.5-0)[86] in R.

Analysis of chromatin states

To characterize the chromatin states of the base positions hit by ASD TAD promoter DNVs in ASD probands (n of DNVs = 2,414) and unaffected siblings (n of DNVs = 1,794), we used the data of reference epigenomes across 127 tissues/cells from the Roadmap Epigenomics Project.[26] We first classified the obtained data into the 17 tissue/cell groups shown in Figure 2A (IMR90, ESC, iPSC, ES-deriv, Blood & T-cell, HSC & B-cell, Mesench, Myosat, Epithelial, Neurosph, Thymus, Brain, Adipose, Muscle, Heart, Sm. Muscle, and Digestive). We then merged the data (bed files) for each of the core 15 chromatin states (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state) in each tissue/cell group using the intersect function of bedtools.[85] This means that a base position that was assigned to a certain chromatin state in any of the data in each tissue/cell group was considered to be in that chromatin state, and a base position can be assigned to multiple chromatin states in a tissue/cell group. Based on these merged bed files for the 17 tissue/cell groups, we annotated if a position of DNV falls into a chromatin state using the intersect function of bedtools.[85] We then analyzed the distribution of ASD gene TAD promoter within a chromatin state and the others in ASD probands and unaffected siblings by two-tailed Fisher's exact test for each chromatin state. Chromatin states of tissue/cell groups for which the number of corresponding ASD gene TAD promoter DNVs (total in cases and controls) was less than or equal to ten were excluded from the analysis. Synthetic P values for each chromatin state were calculated by Fisher's method using the poolr package[87] (v1.1.1) in R. We calculated two sides (i.e. "greater" or "less") of one-tailed P values for each chromatin state and tissue/cell group, combined each side of P values for a chromatin state by Fisher's method, and then used the smaller P value multiplied by two as the significance for the corresponding chromatin state. When combining the P values, we took into account the non-independence of chromatin states of each tissue/cell group and performed adjustments based on the estimates of the effective number of tests. The effective number of tests was obtained by the method described by Nyholt et al.[117] from the correlation matrix of the ORs of ASD gene TAD promoter DNVs for 15 chromatin states in each tissue/cell group using the meff function of poolr.

TF binding enrichment analysis

Analysis of TFs whose binding sites are enriched in the base positions hit by ASD gene TAD promoter DNVs in ASD probands was performed by using the Enrichment Analysis tool of ChIP-Atlas[27] with the following parameters and input files: Cell type Class = Neural, Experiment type = ChIP: TFs and others, Threshold for Significance= 50, Dataset A = bed file of the positions of ASD gene TAD promoter DNVs in ASD probands, Dataset B = bed file of the positions of ASD gene TAD promoter DNVs in unaffected siblings. Statistical significance plotted in the heatmap in Figure S3A and Table S5 was calculated by two-tailed Fisher's exact tests. From the output table, we first extracted experiments where ten or more ASD gene TAD promoter DNVs are located within the TF-bound regions defined in ChIP-Atlas. We then selected results for TFs with five or more experiments and tissues/cells with five or more experiments and used them for the generation of the heatmap in Figure S3A. The synthetic P values for each TF in Figure 2B were obtained by Fisher's method from the one-sided P values, using a method similar to that described in the above section for the analysis of chromatin states.

Prediction scores for non-coding variants

We annotated promoter DNVs with deleteriousness scores for non-coding variants predicted by various tools by using the data compiled in regBase[32] (v1.1.1) based on the positions in the hg19 genome. For comparisons across prediction tools, we used data obtained by the following 22 tools (including different versions of the same tool) for which Phred-scaled scores are available: CADD[28,29] (v1.3, 1.4. and 1.6), CDTS,[33] CScape,[88] CScape_Somatic,[89] DANN,[90] DVAR,[91] Eigen/Eigen_PC,[30] FATHMM (MKL[92] and XF[93]), FIRE,[94] fitCons,[95] FitCons2,[96] FunSeq2,[97] GenoCanyon,[34] LINSIGHT,[31] ncER,[98] Orion,[99] PAFA,[100] and ReMM.[101] Differences between ASD gene TAD promoter DNVs in ASD probands and unaffected siblings were evaluated by the two-tailed Wilcoxon rank sum test.

Annotation of TADs

Using the above-described annotations, we calculated the length of TADs containing a promoter DNV, the distance between a promoter DNV and the nearest ASD gene, the distance between a promoter DNV and the nearest TAD boundary, and the number of TSSs in the TAD containing a promoter DNV. The number of brain enhancers in the TAD containing a promoter DNV was obtained using the data in PsychENCODE (DER-04a_hg38lft_PEC_enhancers). Differences between ASD gene TAD promoter DNVs in ASD probands and unaffected siblings were evaluated by the one-tailed two-sample Kolmogorov-Smirnov test for length/distance and the two-tailed Wilcoxon rank sum test for the number of TSS/enhancer.

Calculation of *de novo* promoter SNV rates for each TAD

We first calculated the rates of *de novo* SNVs hitting any of the promoter regions (upstream 2000 bp of the Gencode v27 genes) in each TAD using the mutational model considering local trinucleotide contexts[35,118] (https://github.com/pjshort/dddMAPS/blob/master/data/forSanger_1KG_mutation_rate_table.txt). Indels were not considered because the mutation rates for these DNVs were not provided in the above resource and their variant call accuracy is in general lower than that of SNVs. By summing up the rates for each TAD, the total *de novo* SNV rate in the promoter regions in any TADs (166.28 Mb in total) per human diploid genome was calculated to be 4.533. This could be considered a reasonable number because it is known that ~1 *de novo* SNVs are observed in ~40Mb exome regions. Based on this theoretical rate, it was expected that there would be ~17,244 *de novo* promoter SNVs in 3,804 children in SSC. However, the observed number (5,340) was much smaller than this expectation. This was mainly because many *de novo* SNVs in the promoter regions did not have an annotation of promoter variants; for example, a DNV hitting the promoter of gene A can be a coding DNV of another gene B, in which case the coding annotation is prioritized. Therefore, we performed an adjustment by multiplying the observed/expected ratio of the numbers of *de novo* promoter SNVs in unaffected siblings of the combined SSC+SPARK cohort by the *de novo* promoter SNV rates for each TAD calculated above. The adjustment factor was calculated as 0.287 = 5336 / (4095 * 4.533), where 4,095 is the total number of unaffected siblings and 5,336 is the total number of observed *de novo* promoter SNVs in 4,095 individuals. The list of expected numbers of *de novo* promoter SNVs for each TAD is provided in Table S6.

TADWAS

Enrichment of *de novo* promoter SNVs in ASD in each TAD was assessed by one-tailed Poisson exact test using the observed and expected counts of *de novo* SNV events. We also performed the same analysis in the unaffected siblings to identify putative hotspots of *de novo* promoter SNV calls, which are less likely to be associated with diseases. We identified a total of 170 TADs enriched for *de novo* promoter SNV at uncorrected $P < 0.05$ in unaffected siblings and excluded them from the analysis. TADome-wide significance was defined as $0.05 / 1826 = 2.74 \times 10^{-5}$ from the number of TADs with one or more *de novo* promoter SNVs in ASD and not enriched for *de novo* promoter SNVs in unaffected siblings. The Manhattan plot in Figure 3A was generated by using the qqman[102] package in R. Full result of TADWAS is provided in Table S6. Hi-C interaction maps of TADs of interest were visualized by the 3D Genome Browser.[36] Genomic loci implicated by both TADWAS and a GWAS meta-analysis[4] were identified by bedtools[85] and visualized by the qqman[102] package in R.

LDSC

We evaluated if the heritability of various traits is enriched in promoter regions of TADs with overrepresentation (uncorrected $P < 0.05$) of promoter DNVs in ASD in the TADWAS using linkage disequilibrium score regression (LDSC)[37,38] (v1.0.1). The promoter regions of each TAD were defined as described in the "Calculation of de novo promoter SNV rates for each TAD" section and liftovered to the hg19 genome. GWAS summary statistics for various psychiatric disorders, including ASD, and traits analyzed in the UK Biobank projects were obtained from the download site of Psychiatric Genomics Consortium[39] (PGC, https://www.med.unc.edu/pgc/download-results/), an ASD GWAS meta-analysis study,[4] and the download site of Alkes Price's Lab (https://alkesgroup.broadinstitute.org/LDSCORE/independent_sumstats/). An analysis of partitioned heritability of the promoter regions in nominally significant TADs (n of TADs = 156) and the promoter regions in the other TADs (n of TAD = 1,670) was performed with the baseline model (v2.2) of 97 genomic annotations and the data of variant frequencies (1000G_Phase3_frq), genotypes (1000G_EUR_Phase3_plink), and weights (1000G_Phase3_weights_hm3_no_MHC) from the1000 Genomes Project Phase3.

Transfection of CRISPR/Cas9 components

Transfection of the components of the CRISPR/Cas9 system and establishment of knock-in iPSCs were performed by referring to the procedures in a previous report.[119] The sequences of single guide RNA (sgRNA) and single-strand oligonucleotides (ssODNs) are shown in Table S9. The sgRNAs were designed by using CHOPCHOP[111] and synthesized by *in vitro* transcription (IVT) using Guide-it™ sgRNA *In vitro* Transcription Kit (Takara Bio Inc.) following the manufacturer's protocol. The ssODNs were generated by Fasmac Co., Ltd. (Kanagawa, Japan). Two days before the transfection, 2.5 × 10^6 cells were plated onto the 10 cm dish coated with iMatrix-511 silk, in the same way as for the passages. On the day of the transfection, the iPSCs on the 10 cm dish were separated by incubation with 2 mL of 1× TrypLE SELECT at 37°C for 5 min. The separated cells were collected with 8 mL of StemFit AK02N medium including 10 μM of Y-27632 after the incubation. The collected cells were centrifuged at 300× g for 3 min, and the pellet was diluted to 2.0 × 10^7 cells/mL by buffer R of Neon™ Transfection System (Thermo Fisher Scientific). The mixture of CRISPR/Cas9 components (24 ng/μL of sgRNA, 0.1 μg/μL of Alt-R S.p. HiFi Cas9 Nuclease V3 [Integrated DNA Technologies, IA, USA], 66.7 ng/μL of pEGFP-N2 plasmid [Clontech Laboratories Inc., CA, USA], and ssODN [the concentration is described below]) was transfected into 1.0 × 10^5 cells by the electroporation using 10 μL scale of Neon™ Transfection System following the manufacturer's protocol. For the knock-in of the promoter variant of interest, two types of ssODNs were prepared. For the variant allele, one includes a selected promoter DNV in ASD and a variant in the PAM sequence to suppress the re-cleavage after the knock-in (mutant-ssODN). The other only possesses the same PAM variant for the generation of the control allele (control-ssODN). The concentrations of ssODNs in the generation of the iPSCs harboring the chr7:g.26200781C>T variant are as follows; 1 μM of the control-ssODN and 1 μM of the mutant-ssODN for the mutant clones, and 2 μM of the control-ssODN for the control lines. The concentrations of ssODNs for the chr22:g.19179942C>T variant are as follows; 2 μM of the control-ssODN and 2 μM of the mutant-ssODN for the mutant clones, and 4 μM of the control-ssODN for the control lines. The condition of the electroporation is 1,300 V, 30 ms, and one pulse. The transfected iPSCs were plated onto one well of the 48 well plates coated by iMatrix-511 silk. The transfection was performed twice for each clone on the same day.

Single-cell sorting

Two days after the transfection, the transfected iPSCs were treated with 100 μL of 1× TrypLE SELECT (Thermo Fisher Scientific) at 37°C for 7 min and collected by 900 μL of the Stemfit AK02N including 10 μM of Y-27632 followed by centrifugation at 300 × g for 3 min at room temperature. The collected cells were suspended by 1 mL of the StemFit AK02N with Y-27632 and stained by 5 μL of 7-AAD (BD Biosciences, NJ, USA) to detect the dead cells. The EGFP positive and 7-AAD negative iPSCs were single-cell-sorted to the iMatrix-511 silk-coated 96 well plates filled with 100 μL of the StemFit AK02N including 10 μM of Y-27632 and 0.2 mg/mL of Primocin(TM) (NACALAI TESQUE, INC.) with BD FACSAria II (BD Biosciences). The cell sorting was performed by the Support Unit for Bio-Material Analysis, Research Resources Division, RIKEN Center for Brain Science.

Screening of knock-in iPSC clones

The day of the single-cell sorting was defined as day 0. On day 3, 100 μL of the culture medium with Y-27632 and Primocin(TM) was added to each well. Half of the medium was exchanged on day 6, and the formation of colonies was identified on day 9. On day 10, the passage of the colonies was performed. The colonies were treated with 30 μL of Accutase (NACALAI TESQUE, INC.) at 37°C for 8 min. Subsequently, 200 μL of StemFit AK02N including 10 μM of Y-27632 was added to the wells treated with Accutase. After the twenty times of pipetting, 100 μL of the suspension of the cells was plated onto 96 (for genotyping) and 12 (for freeze stock) well plates coated by iMatrix-511 silk, respectively. The medium change was conducted every day until the following steps. Genotyping was conducted when the iPSCs plated onto 96 well plates were grown to confluency. The confluent iPSCs were treated with 100 μL of the lysis buffer (10 mM Tris-HCl, 0.1 % Triton-X, 10 mM EDTA, and 1 x proteinase K diluted with sterile water) at 37°C overnight. The proteinase K was inactivated by the incubation at 95°C for 5 min. The PCR reaction for the genotyping was performed by the condition shown in Table S10 using 0.2 μL of the cell lysate derived from each well. The PCR products were sequenced with BigDye Terminator V3.1 on an ABI 3730xl sequencer (Life Technologies, CA, USA). The iPSCs on the 12 well plates with the confirmed introduction of the promoter variant of interest and PAM variants were frozen using STEM CELL BANKER GMP grade (Takara Bio Inc.) following the manufacturer's protocol. A part of the control lines was derived from the iPSCs transfected with both the control-/mutant-ssODN.

Screening of possible off-target variants

Candidate off-target sites were predicted by Cas-OFFinder[112] with the following setting: mismatch = 3 and bulge of DNA/RNA = 2. We did not analyze candidate sites in chromosome Y because the HPS4290 cells are derived from a female human. We considered sites with mismatch <= 3 or mismatch + bulge <= 2 as candidates, and zero and 14 sites meeting these criteria were detected for the sgRNA used to introduce the chr7:g.26200781C>T and the chr22:g.19179942C>T variants, respectively (Table S11). To be further conservative, we also considered candidate sites harboring four mismatches for the sgRNA to introduce the chr7:g.26200781C>T variant and selected five additional candidates by prioritizing those within gene bodies (Table S11). We performed Sanger sequencing of these candidate sites and analyzed if there are off-target variants by comparing the obtained sequences with the reference genome.

RNA-seq

From the iPSCs twice-passaged after thawing the frozen stock, total RNAs were extracted using Trizol reagent (Thermo Fisher Scientific) followed by the treatment with DNaseI (Takara Bio Inc.). The quality of the extracted RNAs was verified by Agilent Bioanalyzer (Agilent Technologies, Ltd., CA, Japan), and samples with RNA integrity number (RIN) >= 8.1 were used for the RNA-seq analysis. The preparation of the library and sequencing were conducted by Novogene Co., Ltd (Beijing, China). The mRNA was purified from the extracted total RNAs using poly-T oligo-attached magnetic beads. After the fragmentation, the first-strand cDNA was synthesized using random hexamer primers, and subsequently, the second-strand cDNA was synthesized, using NEBNext Ultra RNA Library Prep Kit for Illumina (New England BioLabs Inc., MA, USA). The sequencing was performed by Illumina NovaSeq 6000 (Illumina, Inc., CA, USA). The amount of data per sample was ~6Gb (~20M of 150bp × 2 paired-end reads).

Read mapping and generation of count matrices

After the removal of the low-quality reads and adapter sequences by trim_galore[103] (v.0.6.6) with default settings, all filtered reads were mapped on the hg38 genome using STAR[104] (v.2.7.9a) with "–outFilterMultimapNmax 1". The output SAM files were converted to BAM files by Samtools[105] (v.1.3.1). Ribosomal RNAs were removed after the mapping using the intersectBed functions of bedtools[85] (v.2.30.0). The matrices of genes and read counts were generated from the filtered BAM files by using featureCounts[106] (v.2.0.1) with the default setting.

DEG and downstream analysis

Differentially expressed genes (DEGs) were detected by using the DESeq2[107] package (v1.36.0) in R following the RNA-Seq workflow described in ref[120]. To remove unwanted batch effects, we used the sva package[108] (v3.44.0). The number of surrogate variables was determined by using the num.sv function. We extracted results for protein-coding genes defined by BioMart (Ensembl Genes 104, GRCh38.p13) with non-NA P values and used them for downstream analyses. The UCSC Genome Browser (https://genome.ucsc.edu/)[121] and the 3D Genome Browser (http://3dgenome.fsm.northwestern.edu/)[36] (Tissue: H1-NPC, Resolution: 40 kb) were used for the generation of Figures 4B and 4G. Overlap between DEGs and genesets of interest (i.e. SFARI ASD genes and genesets regarding the postmortem brain analyses[58,59]) was assessed by hypergeometric tests. GO enrichment analysis was performed by using Metascape.[109] We selected the top ten significant GO clusters from the results of the Metascape analysis and used them for network visualization by Cytoscape[110] (v3.7.2). In Figures 4E, 4J, and S14E, two nodes were connected when the Kappa similarity is larger than 0.5. Nodes were positioned by the Group Attributes Layout mode based on the cluster IDs with manual adjustments. Node sizes are proportional to the enrichment P values. An additional gene set enrichment analysis was conducted by using GSEA[113] (v.4.3.2). To prepare the input expression dataset for GSEA, we extracted the normalized read counts data from each experimental replicate using the "counts" function of DESeq2[107] after the adjustment for batch effect by sva. We then selected the normalized expression data of the abovementioned protein-coding genes. We used the "c5.go.v2023.1.Hs.symbols.gmt" geneset database with the "Human_Ensembl_Gene_ID_MSigDB.v2023.1.Hs.chip" gene identifiers for the analysis. We used the default settings except that the method parameter of "Number of permutations" was set to 10,000 and the "Permutation type" was set to "gene_set".

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical details are indicated in the STAR Methods and/or in the figure legends. Pearson's correlation coefficient (R) in Figures 4, S18, and S21 was calculated using the "cor.test" function in R.