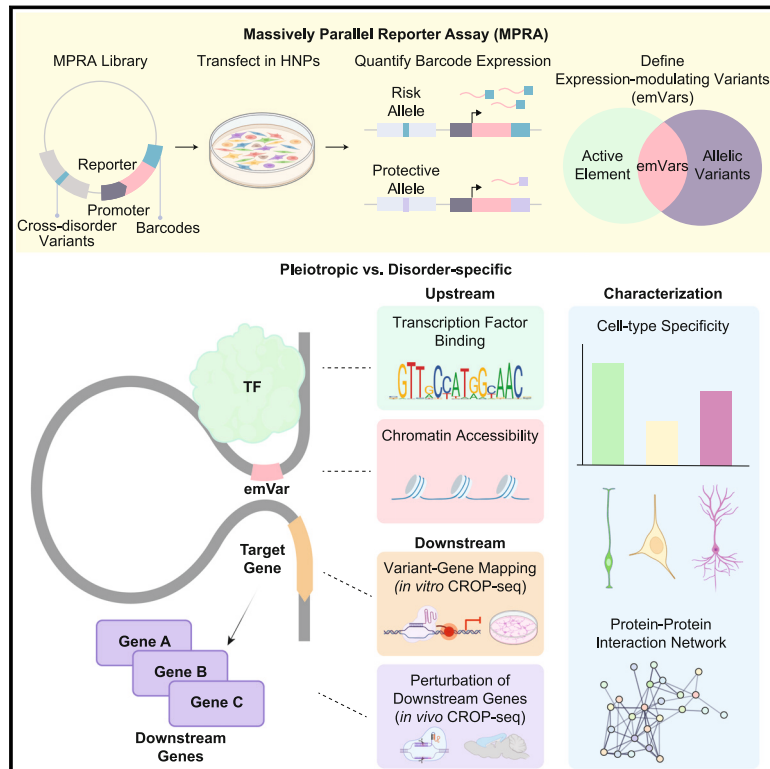# Massively parallel reporter assay investigates shared genetic variants of eight psychiatric disorders

Graphical abstract

Authors

Sool Lee, Jessica C. McAfee, Jiseok Lee, ..., Patrick F. Sullivan, Adriana S. Beltran, Hyejung Won

Correspondence

hyejung_won@med.unc.edu

In brief

High-throughput experimental validation of genetic variants linked to eight psychiatric disorders reveals the regulatory mechanisms underlying variants with pleiotropic and disorder-specific effects.

Highlights

- MPRA tests psychiatric risk variants with pleiotropic and disorder-specific effects

- Pleiotropic variants and genes are active across a broader excitatory neuronal lineage

- Pleiotropic effects are mediated through protein-protein interaction networks

- CRISPR perturbation confirms variant-gene relationships and pleiotropic mechanisms

CelPress

Resource

Massively parallel reporter assay investigates shared genetic variants of eight psychiatric disorders

Sool Lee,[1,2,15] Jessica C. McAfee,[1,2,15] Jiseok Lee,[1,2,15] Alejandro Gomez,[1,2] Austin T. Ledford,[1,2] Declan Clarke,[3,4,5,6,7] Hyunggyu Min,[1,8] Mark B. Gerstein,[3,4,5,6,7] Alan P. Boyle,[9,10] Patrick F. Sullivan,[1,11,12] Adriana S. Beltran,[1,13,14] and Hyejung Won[1,2,16,*]

[1]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2]Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[3]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT 06520, USA
[4]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT 06520, USA
[5]Department of Computer Science, Yale University, New Haven, CT 06520, USA
[6]Department of Statistics & Data Science, Yale University, New Haven, CT 06520, USA
[7]Department of Biomedical Informatics & Data Science, Yale University, New Haven, CT 06520, USA
[8]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[9]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[10]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109, USA
[11]Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27599, USA
[12]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[13]Human Pluripotent Cell Core, University of North Carolina, Chapel Hill, NC 27599, USA
[14]Department of Pharmacology, University of North Carolina, Chapel Hill, NC 27599, USA
[15]These authors contributed equally
[16]Lead contact
*Correspondence: hyejung_won@med.unc.edu
https://doi.org/10.1016/j.cell.2024.12.022

SUMMARY

A meta-genome-wide association study across eight psychiatric disorders has highlighted the genetic architecture of pleiotropy in major psychiatric disorders. However, mechanisms underlying pleiotropic effects of the associated variants remain to be explored. We conducted a massively parallel reporter assay to decode the regulatory logic of variants with pleiotropic and disorder-specific effects. Pleiotropic variants differ from disorder-specific variants by exhibiting chromatin accessibility that extends across diverse cell types in the neuronal lineage and by altering motifs for transcription factors with higher connectivity in protein-protein interaction networks. We mapped pleiotropic and disorder-specific variants to putative target genes using functional genomics approaches and CRISPR perturbation. *In vivo* CRISPR perturbation of a pleiotropic and a disorder-specific gene suggests that pleiotropy may involve the regulation of genes expressed broadly across neuronal cell types and with higher network connectivity.

INTRODUCTION

The World Health Organization estimates that 1 in 8 people (970 million) suffer from a psychiatric disorder of some kind.[1] Because of the high comorbidity between different psychiatric disorders, common genetic etiology has been suggested.[2] A study comparing genetic correlations across 25 different brain disorders suggested shared genetic risk factors (pleiotropy) across psychiatric disorders.[3]

To investigate this further, the Psychiatric Genomics Consortium (PGC)[4] conducted a meta-genome-wide association study (GWAS) of eight psychiatric disorders, including autism spectrum disorder (AUT), attention-deficit/hyperactivity disorder

(ADD), schizophrenia (SCZ), bipolar disorder (BIP), major depressive disorder (MDD), Tourette syndrome (TS), obsessive-compulsive disorder (OCD), and anorexia nervosa (ANO), to identify pleiotropic genetic variants.[5] Among the 136 genome-wide significant (GWS) loci (hereafter referred to as "cross-disorder loci"), 109 loci were associated with more than one disorder, underscoring widespread pleiotropy in psychiatric genetics. Pleiotropy may account for the extensive phenotypic comorbidity observed across different psychiatric disorders.[3,6] Functional characterization of these loci suggested a neurodevelopmental origin,[5] but the mechanisms by which these variants confer broader influence across diagnostic boundaries are unknown.
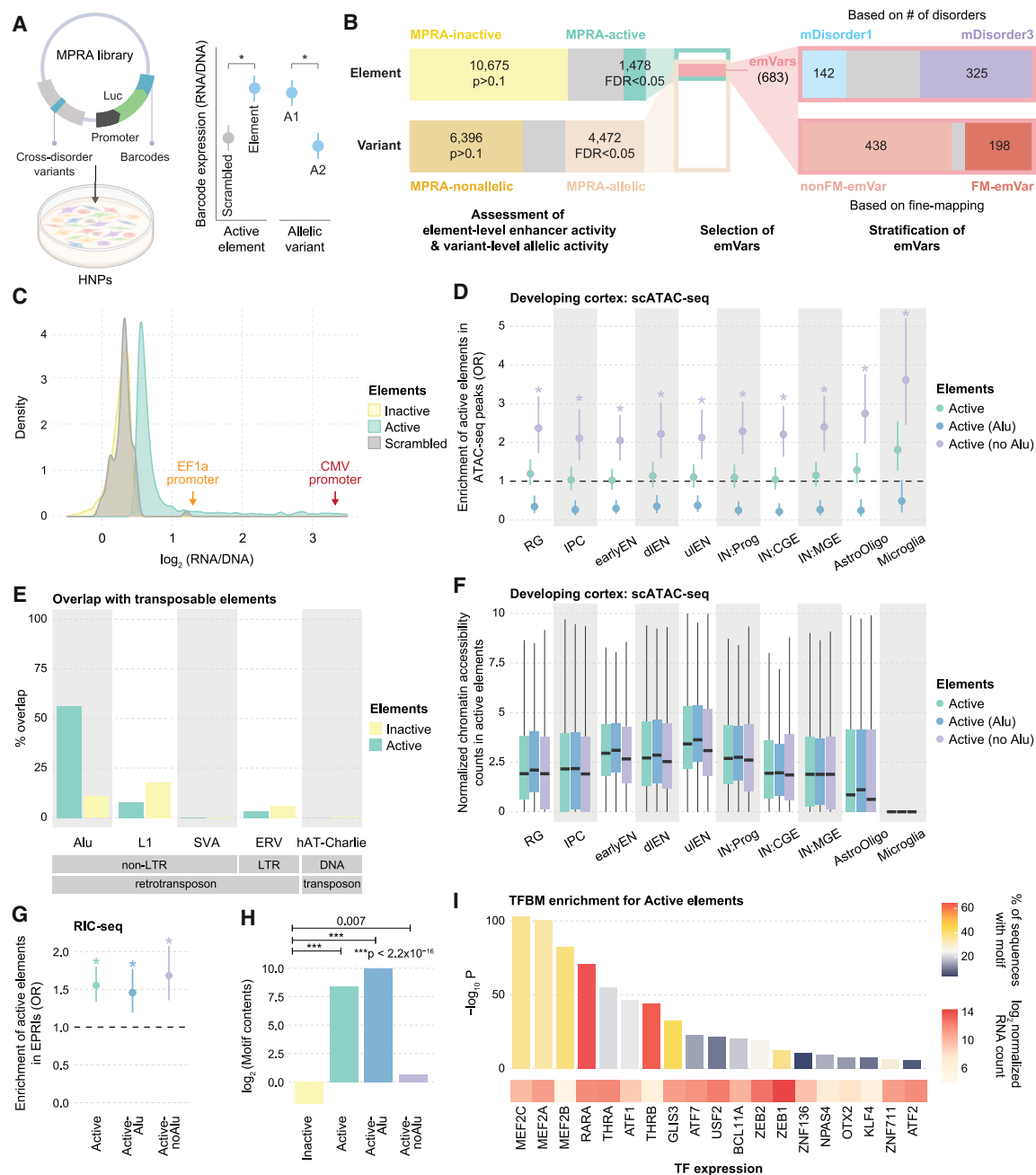
Figure 1. MPRA-active elements are enriched with Alu repeats and TFBMs

(A) We generated a cross-disorder GWAS MPRA library, which was introduced into HNPs. We performed (1) element-level analysis to identify MPRA-active elements with higher enhancer activity than randomly scrambled sequences and (2) variant-level analysis to identify MPRA-allelic variants with differential allelic activity between A1 and A2 alleles.

(B) Summary of the different classes of elements and variants defined in this study. Element-level analysis stratified cross-disorder variants into MPRA-active (false discovery rate using the Benjamini-Hochberg, FDR < 0.05) and MPRA-inactive ($p > 0.1$) elements. Variant-level analysis stratified cross-disorder variants into allelic (FDR < 0.05) and non-allelic ($p > 0.1$) variants. emVars were defined as the intersection of MPRA-active elements and MPRA-allelic variants. emVars were further categorized based on the number of associated psychiatric disorders (mDisorder1 vs. mDisorder3) and computational fine-mapping (fine-mapped [FM] vs. non-fine-mapped [nonFM]).

(C) Enhancer activity (measured by the RNA/DNA ratio of a given element) of MPRA-active and MPRA-inactive elements compared with randomly scrambled sequences. RNA/DNA ratios of positive controls (CMV promoter and EF1α promoter) are also depicted.

(D) Enrichment of MPRA-active, active-Alu, and active-noAlu elements in CREs of different cell types in the developing cortex compared with MPRA-inactive elements. *FDR < 0.05. p values and odds ratio (OR) calculated by two-sided Fisher's exact test. Active-noAlu; two-sided Fisher's exact test, OR > 1, RG FDR = 9.92×10^{-7}, IPC FDR = 1.29×10^{-5}, earlyEN FDR = 4.91×10^{-6}, dlEN FDR = 5.49×10^{-6}, ulEN FDR = 5.49×10^{-6}, IN:Prog FDR = 7.86×10^{-7}, IN:CGE FDR =

(legend continued on next page)

To further examine the potential differences in functional characteristics between variants with pleiotropic effects and those specific to certain disorders, we employed a massively parallel reporter assay (MPRA).[7] Given the evidence supporting a neurodevelopmental origin for these disorders,[5,8–10] we carried out MPRA on 17,841 cross-disorder risk variants (see STAR Methods for selection criteria) in human neural progenitor cells (HNPs) to understand their impact on gene regulation during neurodevelopment. We also included tiled 150 base pair (bp) sections of the cytomegalovirus (CMV) and elongation factor 1 alpha (EF1α) promoters as positive controls and scrambled DNA sequences with matching GC content to the MPRA library as negative controls. The resulting MPRA library consisted of a 150-bp region of the genome with a risk or protective allele in the middle, a minimal promoter, a luciferase reporter gene, and a unique 20-bp barcode. This library was introduced to HNPs, and after 72 h of transfection, DNA and RNA were extracted and subjected to high-throughput barcode sequencing. Allelic expression was summarized by aggregating RNA and DNA barcode counts separately for each allele (Figure S1A). The resulting RNA/DNA ratios, which measure the regulatory activity, demonstrated a high average Pearson correlation coefficient (r = 0.985) across 22 biological replicates (Figure S1B).

RESULTS

Gene regulatory activity of variant-harboring elements

We quantified the enhancer activity of 150-bp variant-harboring elements by comparing them to randomly scrambled sequences (Figure 1A and S1A). Among the 15,902 variant-harboring elements that passed quality control, 1,478 (9.3%) elements showed significantly greater enhancer activity compared with randomly scrambled sequences (MPRA-active elements), while 10,675 (67.1%) elements showed comparable enhancer activity to randomly scrambled sequences (MPRA-inactive elements, Figures 1A–1C and S1C; Table S1, Document S1 available in supplemental information). Positive controls, which are well-established promoters such as the CMV and EF1α promoters, exhibited strong enhancer activity compared with randomly scrambled sequences, confirming the efficacy of MPRA in detecting gene regulatory activity (Figure 1C).

We then compared the overlap of MPRA-active and MPRA-inactive elements with the biochemically defined *cis*-regulatory elements (CREs) of various cell types in the developing brain.[11] Surprisingly, MPRA-active elements were not enriched with brain CREs compared with MPRA-inactive elements, with the exception of microglia (Figure 1D). Because MPRA on GWAS variants associated with type 2 diabetes and blood pressure reported enrichment of MPRA-active elements in Alu repeats,[12,13] we evaluated whether the lack of overlap between MPRA-active elements and CREs could be attributed to transposable elements with poor mappability due to their repetitive nature. Indeed, we found that MPRA-active elements were strongly and selectively enriched with Alu repeats among various types of transposable elements (Figure 1E). Overlap with Alu repeats correlated with the enhancer activity of MPRA-active elements (Figure S1D).

We stratified MPRA-active elements into two categories based on their overlap with Alu repeats: active-Alu, which display a high degree (>50%) of overlap, and active-noAlu, which display low or no overlap (Document S1). Active-noAlu elements were overrepresented with CREs in every cell type of the developing brain, while active-Alu elements were depleted with brain CREs compared with MPRA-inactive elements (Figure 1D). Chromatin accessibility of all three classes of active elements was strongest in upper-layer excitatory neurons (Figure 1F).

A recent study has shown that up to 90% of enhancer-promoter RNA interactions (EPRIs) harbor Alu repeats in the enhancer RNA.[14] Therefore, we assessed whether active-Alu elements are enriched in EPRIs,[14] which effectively capture enhancers with Alu repeats. All active elements, including active-Alu elements, were enriched in EPRIs compared with MPRA-inactive elements (Figure 1G), suggesting that active-Alu elements, which were previously understudied by the traditional definition of enhancers, may play a prominent role in driving gene regulation.

Next, we investigated whether the enhancer activity of MPRA-active elements is mediated by transcription factors (TFs) by quantifying the abundance of TF motifs within a given element (TF motif content scores). MPRA-active elements showed a higher motif content score than MPRA-inactive elements, suggesting that differential MPRA activity on an elemental level is mediated by TF binding (Figure 1H). Furthermore, active-Alu elements exhibited a higher motif content score than active-noAlu elements, suggesting that Alu repeats may drive gene regulation via enhanced TF binding.

We explored TFs with binding motifs (TFBMs) enriched in MPRA-active elements. TFs with robust expression in HNPs

1.62 × 10^{-6}, IN:MGE FDR = 4.85 × 10^{-7}, AstroOligo FDR = 1.49 × 10^{-7}, and Microglia FDR = 1.85 × 10^{-8}. Active-Alu; radial glia (RG) FDR = 0.28, intermediate progenitor cells (IPCs) FDR = 0.80, early excitatory neurons (earlyENs) FDR = 0.85, deep layer excitatory neurons (dlENs) FDR = 0.45, upper-layer excitatory neurons (ulENs) FDR = 0.53, inhibitory neuronal progenitors (IN:Prog) FDR = 0.55, CGE-derived interneurons (IN:CGE) FDR = 0.79, MGE-derived interneurons (IN:MGE) FDR = 0.36, astrocytes and oligodendrocytes (AstroOligo) FDR = 0.14, Microglia FDR = 0.0014. Error bars indicate 95% confidence intervals.

(E) Overlap of MPRA-active and MPRA-inactive elements with different classes of transposable elements. L1, long interspersed nuclear element 1; SVA, SINE-VNTR-Alus; ERV, endogenous retroviruses; LTRs, long terminal repeats.

(F) Chromatin accessibility of MPRA-active, active-Alu, and active-noAlu elements in different cell types in the developing cortex.

(G) Enrichment of MPRA-active, active-Alu, and active-noAlu elements in EPRIs in HNPs. p values and OR calculated by two-sided Fisher's exact test. *FDR < 0.05. Two-sided Fisher's exact test, OR > 1, MPRA-active elements FDR = 3.70 × 10^{-8}, active-Alu elements FDR = 1.91 × 10^{-4}, active-noAlu elements FDR = 3.74 × 10^{-6}. Error bars indicate 95% confidence intervals.

(H) Comparisons of motif content scores between MPRA-inactive, MPRA-active, active-Alu, and active-noAlu elements. p values calculated by two-sided Fisher's exact test. ***p < 2.2 × 10^{-16}.

(I) TFBMs enriched within MPRA-active elements and their respective enrichment, percent of sequence overlap, and their normalized RNA count in HNPs. Cumulative binomial test was used to calculate p values by HOMER.
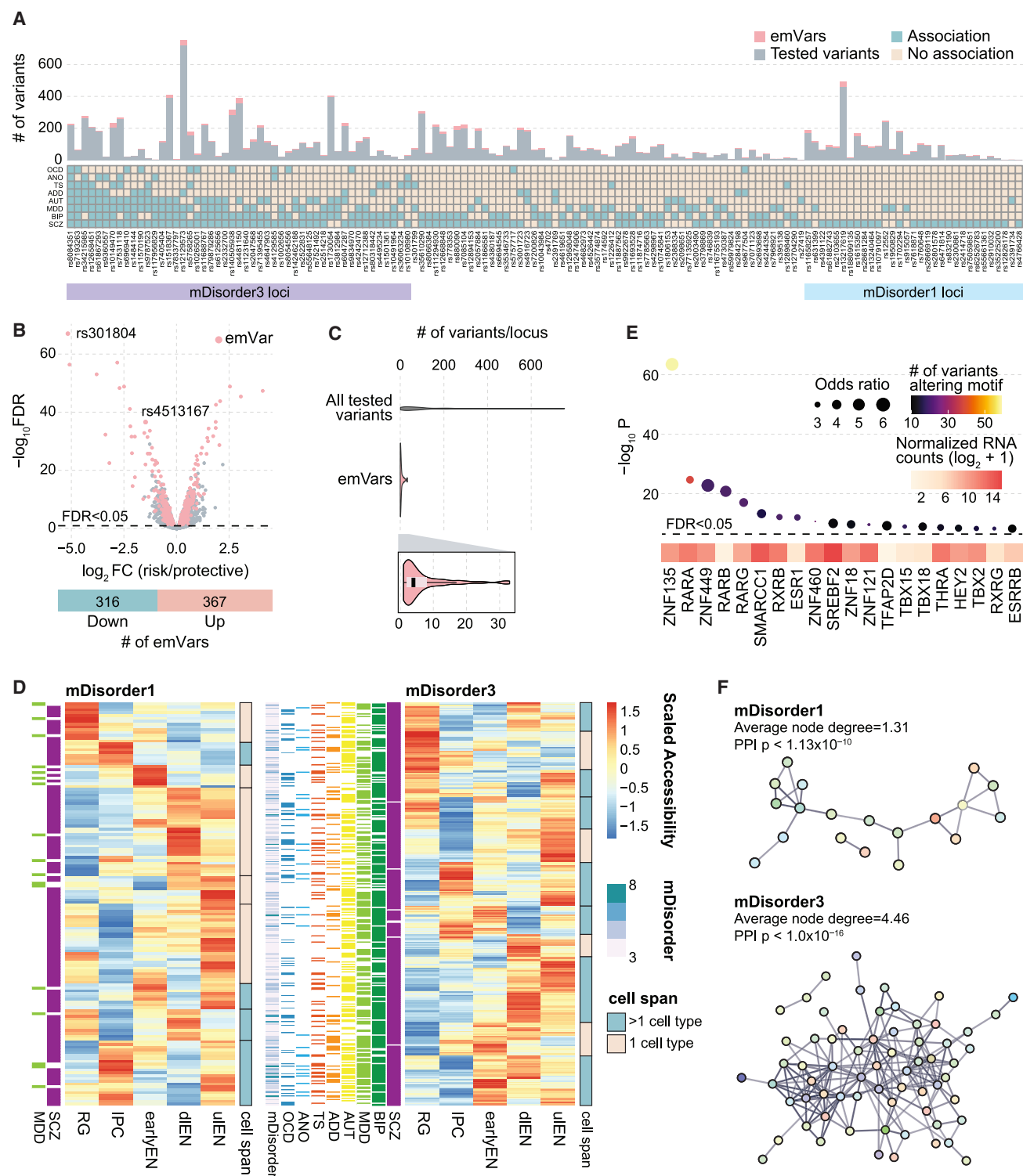
See also Figure S1.

Figure 2. mDisorder3 emVars display chromatin accessibility in broader cell types and alter binding motifs of TFs with higher network connectivity

(A) Top, the number of emVars (pink) out of other tested variants (gray) per locus. Bottom, MPRA-tested cross-disorder GWAS loci as defined by the index SNP and their association with individual psychiatric disorders. If a locus is associated with a disorder (m-value > 0.9 as reported by the original cross-disorder paper[5]), the box is marked cyan; otherwise, it is marked yellow.

(legend continued on next page)

were enriched for MPRA-active elements (Figure 1I). Active-Alu elements were enriched with the same TFs as MPRA-active elements (Figure S1E), indicating that the elevated TF motif content score of MPRA-active elements was largely driven by active-Alu elements (Figure 1H). By contrast, active-noAlu elements were enriched with LEF1, a TF that drives the neuronal proliferation and differentiation,[15] and EGR1, a TF that regulates synaptic plasticity[16] (Figure S1F).

Allelic regulatory activity of psychiatric risk variants

We conducted variant-level analysis to identify variants with differential allelic regulatory effects. Of the 17,841 cross-disorder risk variants tested, 13,311 variants passed quality control (Figure S1A; STAR Methods), spanning all of the tested 136 cross-disorder loci; Figure S2A). Of these 13,311 variants, 683 variants (hereafter referred to as expression-modulating variants or emVars) across 103 loci showed significant allelic regulatory activity within the enhancer context (Figures 1A, 1B, 2A, 2B, S2A, and S2B; Table S1, Document S1). We also defined MPRA-nonallelic variants as those with no significant allelic regulatory activity (Document S1).

On average, we detected 6.6 emVars per locus (median = 4, Figure 2A and 2C). Given that we tested an average of 98 variants per locus (median = 66.5), this result shows that MPRA greatly refines the potentially causal GWAS variants per locus. emVars did not display a strong directional bias, as risk alleles had a ~54% increase and ~46% decrease in expression of the reporter gene (Figure 2B, bottom).

It is sometimes assumed that the GWAS index SNP (the SNP with the smallest p value in a GWS locus) is more likely to be causal compared with the other variants in the locus. To test this assumption, we evaluated the proportion of GWAS index SNPs that were emVars. Out of the 100 GWAS index SNPs tested in our MPRA, only 3 were identified as emVars, recapitulating our previous findings that not all index SNPs exhibit detectable regulatory function in MPRA.[17]

We next surveyed the relationship between GWAS summary statistics and MPRA-measured regulatory activity. Similar to our previous finding,[17] GWAS p values did not differentiate between emVars and MPRA-nonallelic variants (Figure S2C). By contrast, there was a significant difference in GWAS effect sizes ($|GWAS \log_2 OR|$) between emVars and MPRA-nonallelic vari-

ants, suggesting that GWAS variants with larger effect sizes are more likely to have regulatory function (Figure S2D). However, there was no significant correlation between MPRA-allelic activity ($|MPRA \log_2 FC|$) and GWAS effect sizes (Pearson correlation coefficient r = –0.007, p = 0.45).

We further categorized emVars into those located in disorder-specific (mDisorder1: loci/variants are associated with only one disorder) and pleiotropic (mDisorder3: loci/variants are associated with 3 or more disorders) loci (Figures 1B and 2A; Document S1). Pleiotropy did not appear to affect the percentage of emVars, with both mDisorder1 and mDisorder3 loci having ~5% emVars (Figure S2E). Similarly, the number of emVars per locus did not differ between mDisorder1 and mDisorder3 (Figure S2F) but showed a moderate, albeit not significant, increase when the locus was associated with more psychiatric disorders (Figure S2G). The percentage of emVars slightly varied across different disorders, with OCD having the highest percentage (7.1%), while AUT had the lowest percentage (2.7%) (Figure S2H).

Epigenetic characterization of emVars

To gain insight into the epigenetic mechanisms driving variant regulatory activity, we interrogated epigenetic characteristics of emVars. Additionally, we explored a potential difference in epigenetic properties associated with pleiotropy by comparing mDisorder1 emVars with mDisorder3 emVars.

Previous research has shown that many GWAS variants are located within accessible chromatin regions.[18–20] Recapitulating previous research, we observed increased chromatin accessibility in the developing brain for emVars compared with the genomic background SNPs (see STAR Methods for how they were defined; Figure S2I). When examining distinct brain cell types, emVars showed higher chromatin accessibility in the excitatory neuronal lineage, regardless of their pleiotropy (Figure S2I–S2K). We further investigated whether mDisorder1 and mDisorder3 emVars affect different developmental stages across excitatory neuronal lineages. Both mDisorder1 and mDisorder3 emVars showed varied chromatin accessibility within the excitatory neuronal lineage (Figure 2D). However, mDisorder1 emVars displayed more confined chromatin accessibility within a specific cell type, while mDisorder3 emVars had broader accessibility spanning various cell types within the excitatory

(B) Volcano plot displaying allelic regulatory activity of all tested variants. The number of upregulating and downregulating emVars is displayed below in a bar graph. Variants tested by CROP-seq (see Figure 3F and 3H) are labeled.

(C) The number of all MPRA-tested variants (top, gray) and emVars (bottom, pink) per locus. The violin plot of the emVars is zoomed in at the bottom.

(D) Heatmap of chromatin accessibility across excitatory neuronal differentiation for mDisorder1 (left) and mDisorder3 (right) emVars. On the left side of the heatmap, each variant's association with different disorders is indicated by color-coded bars: OCD, dark blue; ANO, light blue; TS, red; ADD, orange; AUT, yellow; MDD, yellow green; BIP, green; SCZ, purple. For mDisorder3 emVars, the mDisorder bar indicates the number of disorders each variant is associated with, with the darkness of the bar representing the extent of association. Heatmaps in the middle depict scaled accessibility for individual mDisorder1 and mDisorder3 emVars across neuronal differentiation stages, including RG, IPC, earlyEN, dlEN, and ulEN. To the right of the heatmap, the cell span for each variant cluster is depicted, indicating whether the variant cluster exhibits chromatin accessibility in only one (1) cell type or across multiple (>1) cell types.

(E) TFs whose motifs are altered by emVars and their expression levels in HNPs. The size of the circles represents OR, indicating the degree to which TFBMs are altered by emVars in comparison to background SNPs. The colors of the circles represent the number of emVars altering the TFBM. Colored squares underneath the plot represent the normalized TF expression level in HNPs.

(F) PPIs of TFs with motifs altered by mDisorder1 (top) and mDisorder3 (bottom) emVars. Average node degrees and p values (PPI p) were obtained from the STRING database. PPI p values depict the enrichment of PPIs compared with a random selection of nodes. Colors are randomly assigned to nodes, which represent TFs.

See also Figures S2 and S3.

neuronal lineage (Figure 2D and S2L), suggesting that mDisorder3 emVars may exert their biological impact across multiple excitatory neuronal lineages.

Genetic variations can alter TFBMs, affecting binding affinity and the expression of associated downstream target genes. To investigate the mechanism by which emVars exert allelic effects, we identified TFBMs that are more likely to be altered by emVars than by genomic background[17] (hereafter referred to as emVar TFs, Table S2). In total, we identified 104 emVar TFs ($p < 0.05$, two-sided Fisher's exact test, 39 TFs with FDR < 0.05), which displayed detectable expression levels in HNPs (Figure 2E). Four emVar TFs (THRA, RARA, TBX21, and P73) overlap with TFs enriched in MPRA-active elements, suggesting that these TFs may be involved in both enhancer and allelic activity. To identify transcriptional regulators associated with pleiotropy in psychiatric disorders, we separately identified TFBMs altered by mDisorder1 and mDisorder3 emVars (hereafter referred to as mDisorder1 TFs and mDisorder3 TFs). We detected 35 mDisorder1 TFs ($p < 0.05$, two-sided Fisher's exact test, 3 TFs with FDR < 0.05, Figure S2M) and 87 mDisorder3 TFs ($p < 0.05$, two-sided Fisher's exact test, 24 with FDR < 0.05, Figure S2N) that are expressed in HNPs.

Since TFs often act cooperatively,[21] we investigated protein-protein interactions (PPIs) between the emVar TFs and observed that mDisorder3 TFs displayed higher protein-protein connectivity than mDisorder1 TFs, indicated by the average node degree (Figure 2F and S2O). This suggests that TFs associated with multiple psychiatric disorders work in concert and cooperate within more interconnected pathways. For example, mDisorder3 TFs include retinoic acid receptors and estrogen receptors (Figure S2N, i.e., RARB, RARG, RARA, RXRB, RXRG, and ESR1), which are responsible for a broad spectrum of neurodevelopmental processes, including neural proliferation, differentiation, regeneration, and synaptic plasticity.[22,23]

Comparisons between statistical and experimental fine-mapping

We compared emVars against probabilistic assessments of variants by statistical fine-mapping algorithms, including FINEMAP,[24] Sum of Single Effects (SuSiE),[25] and CAusal Variants Identification in Associated Regions (CAVIAR)[26] (Table S3). Fine-mapping algorithms are used to calculate model probabilities over a set of causal configurations (i.e., sets of variants sufficient for explaining the observed GWAS summary statistics) in order to prioritize variants for functional validation. However, statistical fine-mapping results do not always align with functional validation results,[17] and it is difficult to pinpoint which fine-mapping algorithm best predicts the causal set due to the lack of benchmark data. We empirically tested all variants with association significance below a certain threshold ($p < 1 \times 10^{-5}$) within a GWS locus, allowing us to systematically compare our results to those from fine-mapping predictions.

We first assessed the concordance between the three fine-mapping algorithms (Figures S3A and S3B). A comparison of credible sets across loci showed that FINEMAP had the smallest number of total variants in the 95% causal confidence sets (with 5,750 variants), whereas CAVIAR had the most (with 7,050 variants). Notably, the three algorithms had a high degree of overlap, such that all FINEMAP-identified variants were also found in

SuSiE, and all additional SuSiE-identified variants were also found in CAVIAR (Figure S3A). We also observed high correlation in the posterior inclusion probabilities (PIPs) from each algorithm (Figure S3B).

We then overlapped emVars with credible sets generated by each fine-mapping algorithm (Figure S3A). We identified 6,805 variants that were in a credible set from FINEMAP, SuSiE, and CAVIAR but were not emVars. These variants were identified as participating in causal configurations by all three fine-mapping algorithms but did not show regulatory activity in our functional assay. We also identified 438 emVars that did not appear in a credible set for any fine-mapping algorithm (nonFM-emVars) and 198 emVars that appeared in a credible set for all three fine-mapping algorithms (FM-emVars, Figure 1B; Document S1). We focus on these two groups in later analyses. Based on these results, we chose to compare our emVars with SuSiE, a state-of-the-art fine-mapping algorithm.[25]

We explored whether PIPs could differentiate between emVars and MPRA-nonallelic variants. SuSiE PIPs showed no significant difference between the two groups (Figure S3C). Since the size and PIP distribution of credible sets vary with the linkage disequilibrium (LD) complexity of GWS loci, we hypothesized that this lack of distinction might be attributed to GWS loci with complex LD, which typically result in larger credible sets with smaller PIPs. Therefore, we stratified GWS loci based on their credible set sizes, where smaller credible set sizes indicate SuSiE's ability to identify fewer putative causal variants with higher PIPs. Comparing the portion of variants exceeding a specific PIP threshold (e.g., PIP > 0.1) between emVars and MPRA-nonallelic variants across different credible set sizes, we found that emVars were more likely to surpass the threshold only in loci with smaller credible sets (Figure S3D). This suggests that both MPRA and statistical fine-mapping yield comparable results in GWS loci with less complex LD where fine-mapping may effectively discern causal variants.

We next examined whether variants that were both fine-mapped and expression-modulating were more likely to confer regulatory function by comparing FM-emVars to nonFM-emVars. There was no difference in significance (measured by p value) and effect size (measured by $\log_2FC$) in MPRA-measured allelic activity between FM-emVars and nonFM-emVars (Figure S3E and S3F). We again applied progressive thresholding based on the size of the credible set and compared the distributions of MPRA-allelic p values and effect sizes between FM-emVars and nonFM-emVars present in each credible set (Figure S3E and S3F). For small credible sets, FM-emVars showed higher significance and effect sizes for allelic regulatory activity (Figure S3E and S3F). The lack of significance in the smallest credible sets can be attributed to the small number of SNPs present in those groups, which limits our power. Overall, these results suggest that statistical fine-mapping is effective for loci with small credible sets and simple LD patterns but may struggle for loci with complex LD and larger credible sets.

emVar overlap with eQTLs

We overlapped emVars with MetaBrain expression quantitative trait loci (hereafter referred to as meta-eQTL, Table S4), a large-scale eQTL resource that encompasses variant-gene expression
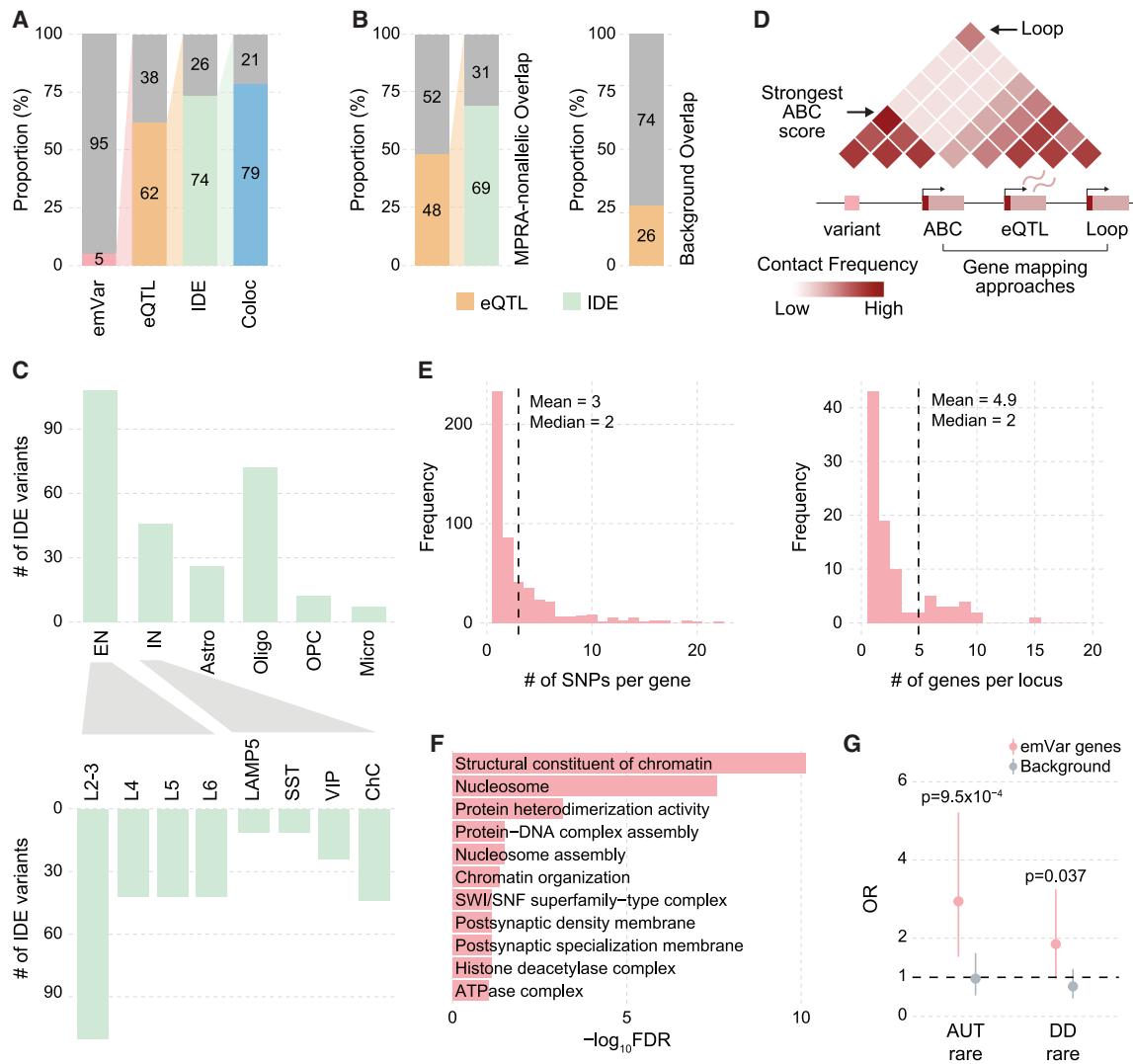
Figure 3. Linking emVars to putative target genes via functional genomics approaches

(A) Overlap between meta-eQTL and emVars. emVars, percent of emVars in total (pink); eQTL, percent of emVars that overlap with eQTLs (orange); IDE, percent of eQTL-overlapping emVars that display identical direction of expression (green); Coloc, percent of loci where IDE variants show colocalization with cross-disorder GWS loci (blue).

(B) Overlap of meta-eQTL with MPRA-nonallelic variants (left) and background SNPs (right).

(C) The number of IDE variants detected in each cell type within the cell-type-specific (top) and neuronal subtype-specific (bottom) eQTL resources. Excitatory neurons (ENs), inhibitory neurons (INs), oligodendrocytes (Oligo), oligodendrocyte precursor cells (OPCs), microglia (Micro), astrocytes (Astro), layer 2–3 excitatory neurons (L2–3), layer 4 neurons (L4), layer 5 neurons (L5), layer 6 neurons (L6), lysosome-associated membrane protein 5-positive interneurons (LAMP5), somatostatin-positive interneurons (SST), vasoactive intestinal polypeptide-positive interneurons (VIP), chandelier cells (ChC).

(D) A schematic showing how emVars were mapped to their putative target genes.

(E) Left, the number of emVars mapped per gene. Right, the number of emVar genes mapped per locus.

(F) GO analysis on emVar genes.

(G) emVar genes are enriched for genes affected by rare PTVs in AUT and DD. Local background genes do not show such enrichment for both AUT and DD risk genes. p values calculated by two-sided Fisher's exact test. Error bars indicate 95% confidence intervals.

See also Figure S3.

relationships from 6,532 human cortical samples.[27] Of the emVars detected in our study, 62% showed eQTL signals in meta-eQTL. Among these overlapping variants, 74% showed an identical direction of effect (IDE) of alleles (i.e., the risk allele is associated with increased gene expression in both MPRA and eQTLs, or vice versa). Of these IDE variants, 79% showed coloc-

alization between cross-disorder GWAS and meta-eQTL (coloc; Figure 3A).

To assess whether the overlap between meta-eQTL and emVars is stronger than what is expected by chance, we measured the eQTL overlap with MPRA-nonallelic and background variants. Of the MPRA-nonallelic variants, 48% showed eQTL

signals, and among these, 69% showed IDE (Figure 3B, left). Of the background variants, only 26% exhibited eQTL signal overlap (Figure 3B, right). Together, emVars display higher eQTL overlap compared with MPRA-nonallelic variants ($p = 3.5 \times 10^{-12}$, two-sided Fisher's exact test) and background variants ($p = 1.4 \times 10^{-87}$, two-sided Fisher's exact test).

Because meta-eQTL are derived from brain homogenate and lack cellular resolution, we also leveraged cell-type-specific eQTL resources to decipher cell-type-specific allelic activity of emVars (Figure 3C). We first compared emVars with cell-type-specific eQTLs from Bryois et al.[28] Among the tested cell types, the largest number of IDE variants was detected in excitatory neurons (ENs; Figure 3C; Table S4), further supporting the epigenetic enrichment of emVars in excitatory neuronal CREs (Figure S2I). Since neurons can be further grouped into neuronal subtypes with different chromatin and transcriptomic architectures,[29–31] we also compared emVars with neuronal subtype-specific eQTLs from the PsychENCODE consortium[32] (Figure 3C). The largest number of IDE variants was detected in layer 2–3 ENs (L2–3), consistent with our findings that emVars have the strongest chromatin accessibility in upper-layer ENs (Figure S2I).

We next evaluated whether pleiotropic variants differ from disorder-specific variants with respect to their overlap with eQTLs. We compared mDisorder1 and mDisorder3 emVars with meta-eQTL,[27] finding that they did not substantially differ in terms of eQTL overlap (67% and 64%, respectively) or percentage of variants that are both IDE and coloc (70% and 65%, respectively, Figure S4A). We also tested the hypothesis that pleiotropic variants may exert their regulatory effects via a greater number of eGenes or across a broader range of cell types. The number of eGenes linked to mDisorder1 and mDisorder3 emVars did not significantly differ for each eQTL dataset (Figure S4B), indicating that pleiotropic effects of variants are not mediated by the number of genes for which expression levels are associated with the variants. Also, in both cell-type-specific eQTL datasets, there was no significant difference in the number of cell types in which eQTLs shared IDE between mDisorder1 and mDisorder3 emVars (Figure S4C). In summary, mDisorder1 and mDisorder3 emVars do not differ by the eQTL overlap, the number of associated eGenes, and the cell types in which eQTLs are observed.

Linking emVars to their putative target genes

To further understand the functional consequence of emVars, we assigned them to their candidate target genes, referred to as emVar genes, using a three-tiered approach (Figure 3D). First, given the substantial overlap between emVars and meta-eQTL (Figure 3A), we used eGenes associated with emVars. Second, we mapped emVars to their respective target genes, building upon our previous study that demonstrated the ability of the activity-by-contact (ABC) model to connect MPRA-validated variants to their putative target genes.[17] Third, we leveraged long-range chromatin interactions, since the ABC model prioritizes proximal target genes. In total, we identified 488 emVar genes (Table S4). On average, ~3 SNPs were mapped to each gene (Figure 3E, left), and ~5 genes were mapped per locus (Figure 3E, right). Similar to our observations with eGenes (Figure S4B), the number of emVar genes did not differ between mDisorder1 and

mDisorder3 loci across different gene mapping approaches (Figure S4D).

We next performed Gene Ontology (GO) analysis using emVar genes to explore their biological function. emVar genes were enriched for genes that encode synaptic proteins and transcriptional regulators, which is consistent with the growing body of evidence that those biological pathways are involved with various psychiatric disorders[8,33–35] (Figure 3F). To mitigate a potential bias in GO enrichment resulting from the biological properties of cross-disorder loci, we defined local background genes as those located nearby cross-disorder loci. While these local background genes showed comparable expression levels to emVar genes in HNPs (Figure S4E), their GO enrichment differed vastly from that of emVar genes (Figure S4F).

In addition, emVar genes were significantly enriched for genes that harbor an excess of rare protein-truncating variants (PTVs) in individuals with AUT[36] and developmental disorders (DDs)[37] (Figure 3G). This finding strengthens the previously reported molecular convergence between rare and common genetic variations in neurodevelopmental and psychiatric disorders.[8,38] This enrichment was not due to the characteristics of genes nearby cross-disorder loci, as we did not observe such enrichment using local background genes (Figure 3G).

Validation of emVar-gene relationships using CRISPR perturbation

Next, we sought to functionally validate the predictive power of the ABC model for establishing variant-gene relationships. To functionally validate emVar-gene relationships, we used CRISPR droplet sequencing (CROP-seq)[39] in human induced pluripotent stem cell (hiPSC)-derived neurons that constitutively express dCas9-KRAB and an inducible *NGN2* cassette.[40] CROP-seq allows simultaneous perturbation of multiple variants at single-cell resolution in a high-throughput fashion (Figure 4A). We generated a homogeneous population of mature glutamatergic neurons, evidenced by transcriptomic profiling of single-cell RNA sequencing (scRNA-seq) data as well as immunofluorescence staining (Figure S5A and S5B).

We selected an emVar, rs301804, as it exhibited the highest significance (allelic FDR = 9.07×10^{-68}) and effect size (allelic log$_2$FC = −5.17) across all variants tested in our MPRA (Figure 2B). This variant is located in the *RERE* locus (Figure 4B) and is associated with two neuropsychiatric disorders (MDD and TS). The ABC model linked rs301804 to *RERE*, which is a gene implicated in positively regulating retinoic acid signaling and linked to neurodevelopmental defects, hypotonia, seizures, and diverse organ anomalies.[41]

CROP-seq targeting of rs301804 led to significant downregulation of *RERE* (negative binomial generalized linear model [GLM] coefficient = −0.129; FDR = 2.42×10^{-5}), while other surrounding genes were not significantly affected (Figure 4C; Table S5). This is intriguing because the nearest promoter to rs301804 is not *RERE* but *SLC45A1*, underscoring the importance of predictive models in elucidating complex gene regulatory networks.

Additionally, we selected the *DCC* locus because it was featured as having the highest pleiotropic association across all eight psychiatric disorders[5] (Figure 4D). Within this locus, the emVar rs4513167 exhibited the highest significance (allelic
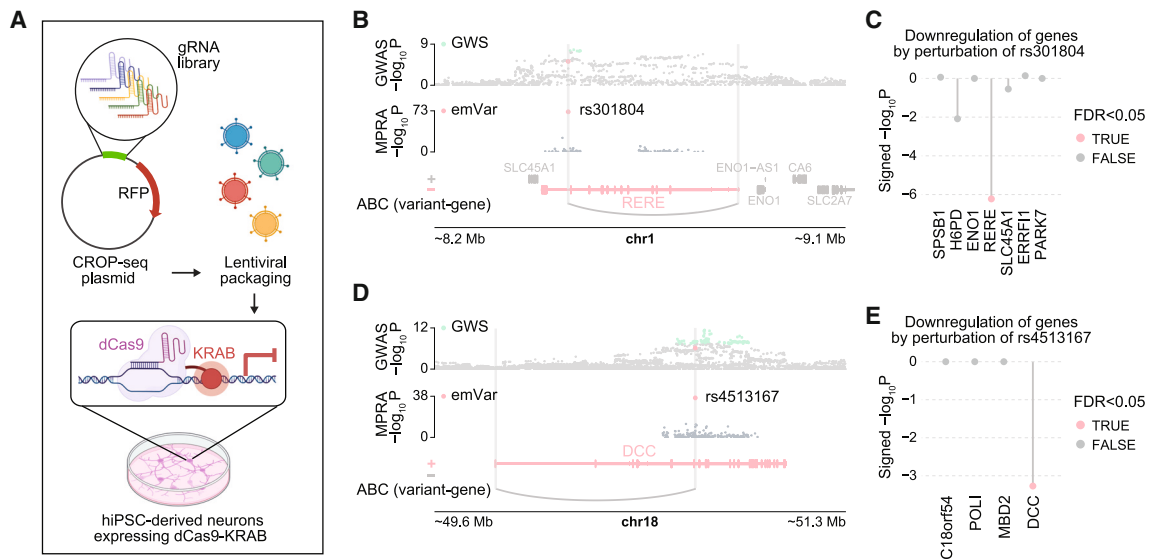
Figure 4. CRISPR validation of variant-gene relationships

(A) A schematic depicting the process of CROP-seq gRNA library generation, lentiviral packaging, and delivery to hiPSC-derived neurons constitutively expressing dCas9-KRAB.

(B) Locus plot for emVar rs301804 in the *RERE* locus. Top panel shows GWAS association statistics of the emVars (pink) and GWS variants (lime green). Middle panel shows significance of MPRA-allelic activity of rs301804 (pink). Bottom panel shows genes within a 900 kb window around the target variant, with *RERE* (pink) as a predicted target gene based on the ABC model (variant-gene relationship displayed with a gray loop). Gray bars highlight variant and gene promoter locations.

(C) Lollipop plot shows perturbation at the rs301804 target site was specific to *RERE* in comparison to nearby genes. Signed $-\log_{10}p$ [Signed $-\log_{10}p = -\log_{10} p_{\text{left-tailed}} \times \text{sign}(\text{GLM}_{\text{coef}})$], where $p_{\text{left-tailed}}$ and GLM_{coef} represent the left-tailed p value and GLM coefficient of a given variant-gene pair, respectively, was plotted for each gene target in the *RERE* locus to represent both the direction of the effect as well as significance. The gene targets meeting FDR < 0.05 are denoted in pink, while those that do not meet significance are denoted in gray.

(D) Locus plot for emVar rs4513167 in the *DCC* locus. Top panel shows GWAS association statistics of the emVars (pink) and GWS variants (lime green). Middle panel shows significance of MPRA-allelic activity of rs4513167 (pink). Bottom panel shows genes within a 1.7 Mb window around the target variant, with *DCC* (pink) as a predicted target gene based on the ABC model (variant-gene relationship displayed with a gray loop). Gray bars highlight variant and gene promoter locations.

(E) Lollipop plot shows perturbation at the rs4513167 target site was specific to *DCC* in comparison to nearby genes. Signed $-\log_{10}p$ was plotted for each gene target in the *DCC* locus to represent both the direction of the effect as well as significance. The gene targets meeting FDR < 0.05 are denoted in pink, while those that do not meet significance are denoted in gray.

See also Figure S5.

FDR = 2.43×10^{-34}) and effect size (allelic $\log_2$FC = −1.27) and was thus selected for functional validation (Figure 2B). The ABC model also suggested a relationship between rs4513167 and *DCC*, which encodes a receptor for Netrin-1, known for its involvement in axon guidance and neurodevelopment.[42] Introducing guide RNAs (gRNAs) targeting the rs4513167 region resulted in selective and significant downregulation of *DCC* (GLM coefficient = −0.261; FDR = 1.21×10^{-2}), while other surrounding genes were not significantly affected (Figure 4E; Table S5).

Compared with the *RERE* locus, the *DCC* locus is relatively gene-poor. *DCC* spans over 1 Mb, and the second-nearest gene to rs4513167 lies ~990 kb away. To assess the specificity of the association between rs4513167 and *DCC*, we conducted functional validation on three additional variants (rs8089270, rs6508210, and rs4614799) in the locus. Among these three variants, rs4614799 was an emVar with the second-highest MPRA-allelic ($\log_2$FC = −0.85) and element ($\log_2$FC = 0.50) effect sizes in the locus (Figure S5C). By contrast, rs8089270 and rs6508210 exhibited differential allelic activity (rs8089270, FDR = 1.15×10^{-2}; rs6508210, FDR = 1.62×10^{-3}), but not element-level

enhancer activity (rs8089270, FDR = 0.903; rs6508210, FDR = 0.394). Notably, CRISPR-mediated perturbation of rs4614799 (emVar) resulted in significant downregulation of *DCC* (FDR = 0.040, GLM coefficient = −0.16), whereas neither rs8089270 nor rs6508210 significantly affected *DCC* expression (Figure S5D; Table S5). Again, the observed regulatory effects of variants could not be solely attributed to distance, as rs8089270 was the closest to the DCC promoter. Instead, the regulatory effect sizes of variants corresponded to the element-level enhancer activity, as measured by MPRA (Figure S5E). These findings suggest that not all variants within the *DCC* locus have gene regulatory functions, emphasizing the importance of modeling variant effects on gene regulation by incorporating both variant-level allelic activity and element-level enhancer activity.

Differences between pleiotropic and disorder-specific genes

We investigated whether emVar genes are engaged in distinct biological processes between pleiotropic and disorder-specific

loci by comparing mDisorder1 genes (143 genes mapped to mDisorder1 emVars) against mDisorder3 genes (182 genes mapped to mDisorder3 emVars, Table S4). mDisorder1 genes were enriched for biological processes encompassing histone deacetylation, pyrimidine metabolism, telomere function, and cell cycles (Figure S6A, top). mDisorder3 genes were enriched for biological processes that involve FGFR1 signaling (Figure S6A, bottom).

Given that mDisorder1 and mDisorder3 emVars showed distinct chromatin accessibility signatures across the neural lineage (Figure 2D), we evaluated whether their target genes are also engaged in different stages of neural differentiation. Consistent with what was observed at the variant level, mDisorder1 gene expression was more likely to be restricted within a specific cell type, while mDisorder3 gene expression spanned multiple cell types during neural differentiation (Figures 5A and 5B). Because distinct developmental expression trajectories between genes mapped to mDisorder1 and mDisorder3 loci were previously reported,[5] we examined cortical expression profiles[43] of mDisorder1 and mDisorder3 genes across developmental epochs. We recapitulated the previous finding that pleiotropic genes show peak expression during mid-gestation, while disorder-specific genes show peak expression during the first trimester[5] (Figure S6B).

We next compared mutation intolerance (measured by loss-of-function observed/expected upper bound fraction [LOEUF] score) between mDisorder1 and mDisorder3 genes. mDisorder3 genes showed higher mutation intolerance compared with mDisorder1 genes, suggesting that pleiotropic variants may influence regulation of mutation-intolerant genes (Figure 5C).

Finally, we evaluated the hypothesis that hub genes, which exhibit high connectivity with other genes in either co-expression or PPI networks, can account for pleiotropy by participating in diverse biological functions through their interaction partners within the network. We first compared kME values, which represent the centrality of a gene within its associated co-expression network,[33] between mDisorder1 and mDisorder3 genes. We did not observe any significant difference between the two groups (Figure S6C). We next examined the connectivity of proteins encoded by mDisorder1 and mDisorder3 genes in the PPI networks. Notably, proteins encoded by mDisorder3 genes demonstrated higher connectivity than proteins encoded by mDisorder1 genes in PPI networks (Figures 5D and 5E).

CRISPR perturbation of pleiotropic vs. disorder-specific genes

To further explore molecular mechanisms underlying pleiotropy, we employed CROP-seq on two genes mapped to either a pleiotropic (ANP32E) or a disorder-specific (KMT5A) locus (Figure 6A). ANP32E is mapped to emVars via the ABC model, and there were no ANP32E eQTLs overlapping with emVars (Figure 6B, left). Meanwhile, KMT5A is mapped to emVars via both meta-eQTL and the ABC model (Figure 6B, right). ANP32E encodes a histone chaperone protein that binds to the histone variant H2A.Z and suppresses the accumulation of H2A.Z.[44] Suppression of H2A.Z accumulation reduces chromatin accessibility, thereby making ANP32E a negative regulator of chromatin accessibility.[44] KMT5A (also known as SETD8) encodes a his-

tone methyltransferase that methylates the lysine-20 of histone H4 (H4K20me).[45] H4K20me has been linked to both transcriptional activation and repression.[46]

We sought to determine which downstream targets, cell types, and networks are affected by perturbing these two transcriptional regulators. We inserted gRNAs targeting the coding sequence of mouse Anp32e and Kmt5a (gRNA sequences available in Table S6) into the lentiviral CROP-seq vector[39] and injected the gRNA lenti-library into the lateral ventricles of the brains of neonatal Cas9 transgenic mice[47] (Figure 6A). After two weeks, we dissociated single cells from the neocortex, sorted for cells expressing the CROP-seq virus using fluorescence-activated cell sorting (FACS), and performed scRNA-seq (Figure 6A). A total of 83,386 neocortical cells from ten experimental batches were used for differential expression analysis. Anp32e and Kmt5a perturbation derived 46 and 244 differentially expressed genes (DEGs), respectively (Table S5), whose overlap was minimal (Figure 6C).

Anp32e DEGs were most highly expressed in both excitatory and inhibitory neurons in the mouse brain (Figure 6D, middle top), and corresponding human ortholog genes were also highly expressed in ENs in the human fetal brain[48] (Figure 6D, middle bottom). By contrast, Kmt5a DEGs were most highly expressed in glial cells in both mouse (Figure 6D, right top) and human fetal brains[48] (Figure 6D, right bottom). In accordance with neuron-biased expression, Anp32e DEGs were enriched in GO terms related to protein complexes involved in synaptic function (Figure 6E, top), while glia-enriched Kmt5a DEGs were annotated with GO terms related to cilium (Figure 6E, bottom).

We also examined whether Anp32e and Kmt5a DEGs were dysregulated in the postmortem brains of individuals with SCZ[49] (Figure 6F). Human orthologs of Anp32e DEGs showed reduced expression in multiple neuronal subtypes in SCZ samples (Figure 6F), whereas Kmt5a DEGs showed increased expression in a smaller subset of cell types, including neurons and astrocytes (Figure 6F). Lastly, Anp32e DEGs exhibited higher node degree (Figure 6G) and connectivity (Figure 6H) in PPI networks compared with Kmt5a DEGs, even though the number of DEGs is larger for Kmt5a perturbation (Figure 6C). This is consistent with our observation that the number of mapped target genes does not explain pleiotropy (Figure S4D).

Overall, our analyses of transcriptomic architecture upon CRISPR perturbation suggest a mechanism by which a pleiotropic gene's effect may propagate through complex PPI networks and expression across diverse neuronal cells. This is consistent with the finding that mDisorder3 variants show chromatin accessibility across broader excitatory neuronal lineages (Figure 2D) and that mDisorder3 genes exhibit higher network connectivity (Figure 5D and 5E).

DISCUSSION

We functionally validated cross-disorder GWAS loci, identifying active elements and variants with gene regulatory activity. We assigned emVars to their putative target genes using eQTLs and chromatin architecture. The relationships between a few emVars and their predicted target genes were further confirmed via CROP-seq in hiPSC-derived neurons. We found that
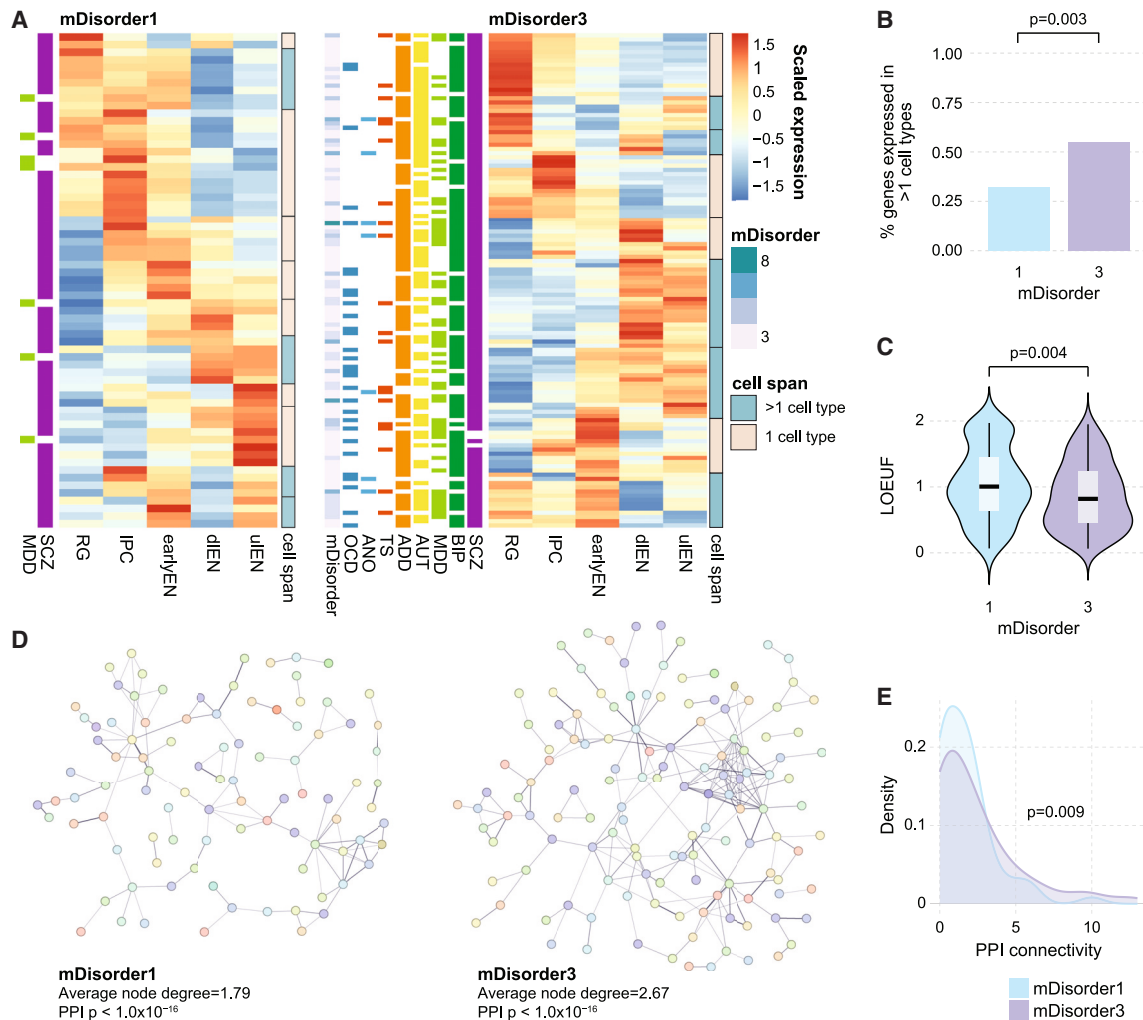
Figure 5. mDisorder1 and mDisorder3 genes differ in expression patterns and network connectivity

(A) mDisorder1 (left) and mDisorder3 (right) gene expression across neuronal differentiation. On the left side of the heatmap, each gene's association with different disorders is indicated by color-coded bars: OCD, dark blue; ANO, light blue; TS, red; ADD, orange; AUT, yellow; MDD, yellow green; BIP, green; SCZ, purple. For mDisorder3 genes, the mDisorder bar indicates the number of disorders with which each gene is associated, with the darkness of the bar representing the extent of association. Heatmaps in the middle depict scaled expression values for individual mDisorder1 and mDisorder3 genes across neuronal differentiation stages. To the right of the heatmap, the cell span for each gene cluster is depicted, indicating whether the gene cluster is expressed in only one (1) cell type or across multiple (>1) cell types.

(B) Expression of mDisorder3 genes is more likely to encompass multiple developmental stages in comparison to mDisorder1 genes, $p = 0.003$. p value calculated by two-sided Fisher's exact test.

(C) Distribution of LOEUF scores for mDisorder1 and mDisorder3 genes, $p = 0.004$. p value calculated by two-sided Wilcoxon rank sum test.

(D) Visualization of PPI networks for mDisorder1 (left) and mDisorder3 (right) genes. Average node degrees and p values were obtained from the STRING database. PPI p values depict the enrichment of PPIs compared with a random selection of nodes. Colors are randomly assigned to nodes, which represent mDisorder1 (left) and mDisorder 3 (right) genes.

(E) mDisorder3 genes are more highly connected in the PPI networks compared with mDisorder1 genes, $p = 0.009$. p value calculated by Anderson-Darling k-Sample test.

See also Figure S6.

upstream (TFs that bind to emVars) and downstream regulators (putative target genes) of pleiotropic variants display higher connectivity in PPI networks. Moreover, pleiotropic emVars and their putative target genes demonstrated broader chromatin accessibility and expression across excitatory neuronal lineages, respectively. Finally, using CROP-seq in the mouse brain, we found that perturbation of a pleiotropic gene can result in

DEGs with neuronal expression and a more interconnected PPI network compared with the DEGs of a disorder-specific gene. Together, understanding the properties and target genes of pleiotropic variants unveils regulatory principles and molecular mechanisms that may contribute to the shared comorbidities in psychiatric disorders. We propose that pleiotropy in psychiatric disorders may be mediated by three mechanisms: regulation
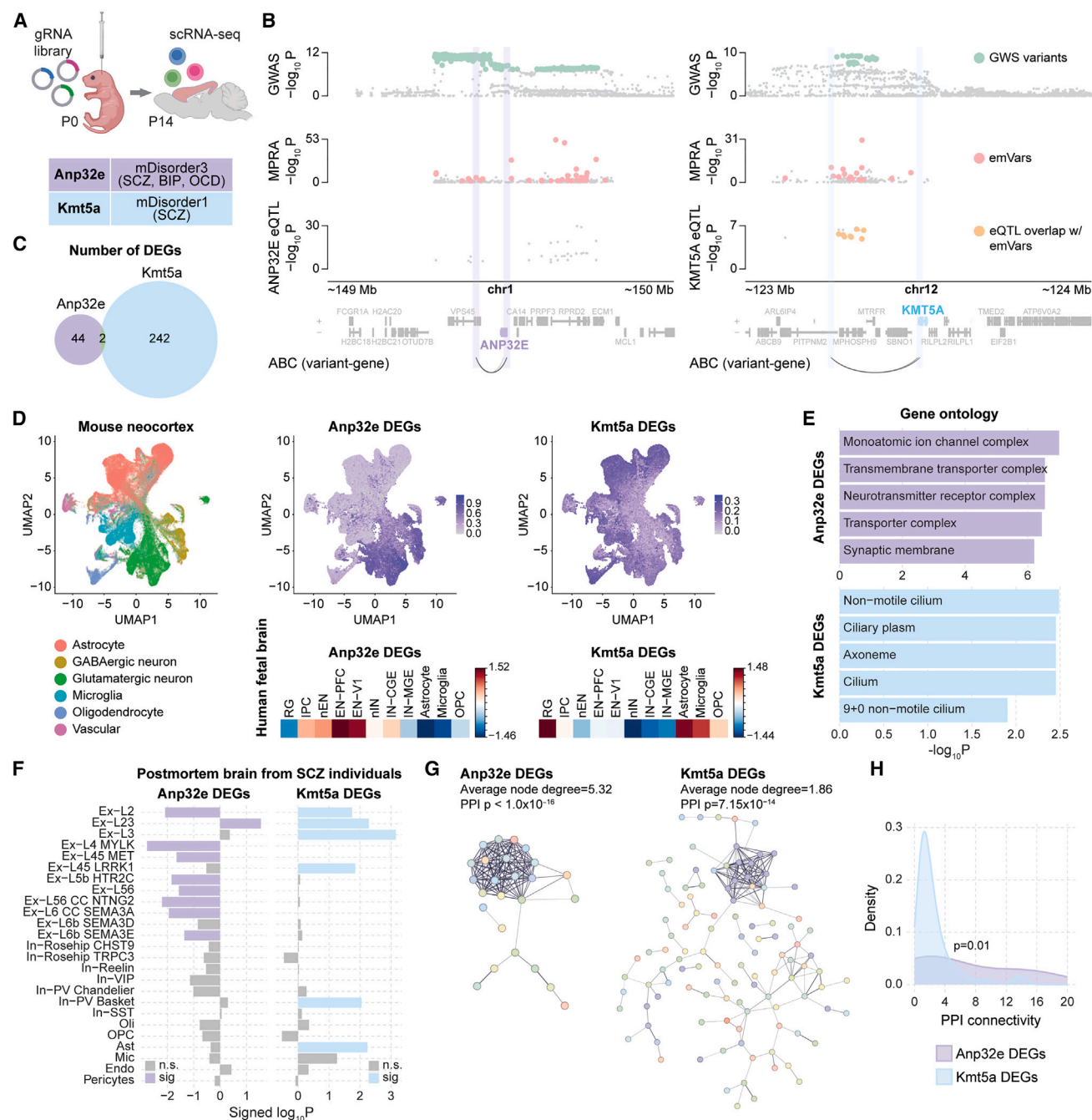
Figure 6. CRISPR perturbations of *Anp32e* and *Kmt5a* lead to dysregulation of genes with different cell type expression and network connectivity

(A) A CRISPR gRNA library targeting an mDisorder3 gene, *Anp32e*, and an mDisorder1 gene, *Kmt5a*, is injected into postnatal day 0 (P0) Cas9 transgenic mice. At P14, neocortical cells are dissociated for scRNA-seq.

(B) Locus plots for *ANP32E* and *KMT5A* showing GWS variants, emVars, eQTLs overlapping with emVars, gene tracks, and variant-gene connections mapped by the ABC model.

(C) Number of DEGs for *Anp32e* and *Kmt5a* CRISPR perturbations.

(D) Left, a uniform manifold approximation and projection (UMAP) plot displays the six major cell types in the mouse neocortex derived from CROP-seq. Middle, *Anp32e* DEGs are highly expressed in neuronal cells in both mouse (top) and human fetal brain (bottom). Right, *Kmt5a* DEGs are highly expressed in glial cells in both mouse (top) and human fetal brain (bottom). Radial glia (RG), intermediate progenitor cells (IPCs), neonatal excitatory neuron (nEN), excitatory neuron in the prefrontal cortex (EN-PFC), excitatory neuron in the visual cortex (EN-V1), neonatal inhibitory neuron (nIN), inhibitory neuron in the caudal ganglionic eminence (IN-CGE), inhibitory neuron in the medial ganglionic eminence (IN-MGE), oligodendrocyte precursor cell (OPC).

(legend continued on next page)

of genes affecting multiple stages of neural differentiation, higher mutational constraint, and higher connectivity within the PPI network. Gaining insights into the shared genetic underpinnings of psychiatric disorders may offer potential therapeutic avenues that can alleviate symptoms for multiple psychiatric disorders.

Limitations of the study

Because we conducted MPRA in an episomal construct, variant regulatory activity is unaffected by chromatin architecture or by local genomic regions beyond the tested 150-bp. Furthermore, our MPRA design mainly prioritizes variants with allelic enhancer activity and may not capture variants with different regulatory effects, such as variants involved in alternative splicing. MPRA assesses the transcription of a reporter gene and barcode but does not indicate the target gene(s) affected by emVars. To establish these connections, we combined MPRA with orthogonal data types, such as eQTLs, Hi-C, and CRISPR-based perturbations. We acknowledge that our CRISPR-based perturbation is limited to a few variants and target genes. Further investigation is warranted to comprehensively identify variant-gene relationships and potential mechanisms by which variants exert pleiotropic effects. It is also important to note that mDisorder1 variants were composed mainly of SCZ variants (~81.7%) due to the higher statistical power of SCZ GWAS. Larger GWAS across different psychiatric disorders will enable us to determine whether the properties of mDisorder1 variants can be extrapolated to other disorder-specific variants. Lastly, while our results suggest that emVars are likely to include causal variants, it does not mean that all emVars are truly "causal" for the disease. Endogenous genome editing, combined with the analyses of downstream biological and clinical phenotypes, would be required to determine the true causal relationship between a variant and the disease.

RESOURCE AVAILABILITY

Lead contact

Any inquiries about the project, analytical results, or other information should be directed to Hyejung Won (hyejung_won@med.unc.edu).

Materials availability

The modified CROP-seq vector is deposited in Addgene (Addgene: ID#230938).

Data and code availability

MPRA sequencing and processed data are available on the Gene Expression Omnibus (GEO) under accession number GEO: GSE244011. Raw and processed scRNA-seq data from CROP-seq on hiPSC-derived neurons and mice are available under GEO accession numbers GEO: GSE276947 and GEO: GSE282731, respectively. Custom codes used to analyze MPRA data are available on our GitHub page (https://github.com/thewonlab/crossdisorder-MPRA).

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Cross-disorder variant selection
  - MPRA library construction
  - Introduction of MPRA libraries into HNPs
  - RNA and DNA library generation
  - CROP-seq in hiPSC-derived neurons
  - CROP-seq in hiPSC-derived neurons
  - Immunofluorescence staining of hiPSC-derived neurons
  - CROP-seq in mouse brain
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - MPRA data preprocessing

(E) Top, top GO terms for *Anp32e* DEGs are related to synaptic functions. Bottom, top GO terms for *Kmt5a* DEGs are related to cilium.

(F) Left, *Anp32e* DEGs are downregulated in a broad range of neuronal cell types in the postmortem brains of individuals with schizophrenia. Right, *Kmt5a* DEGs are upregulated in a smaller set of cell types.

(G) *Anp32e* DEGs show higher node degree in PPI network compared with *Kmt5a* DEGs. Average node degrees and p values were obtained from the STRING database. PPI p values depict the enrichment of PPIs compared with a random selection of nodes.

(H) *Anp32e* DEGs show higher PPI connectivity (number of connected edges per node) than *Kmt5a* DEGs. p value calculated by Anderson-Darling k-Sample test.

- ○ Identification of active elements
- ○ Stratification of active elements based on transposable elements
- ○ Epigenetic annotation of active elements
- ○ TF motif enrichment analysis
- ○ Identification of regulatory variants
- ○ Stratification of variants based on mDisorder
- ○ Epigenetic annotation of emVars
- ○ Defining genomic background SNPs
- ○ TFBM alteration enrichment analysis
- ○ PPI network of TFBMs
- ○ Statistical fine-mapping of emVars
- ○ eQTL overlap
- ○ Gene mapping for emVars
- ○ Stratification of genes based on mDisorder
- ○ Definition of local background genes
- ○ Biological pathway analysis
- ○ Comparison between rare and common genes
- ○ Chromatin accessibility and gene expression profiling across neurodevelopment
- ○ Co-expression and PPI network analysis
- ○ Developmental expression trajectory
- ○ CROP-seq data analysis in hiPSC-derived neurons
- ○ Mouse CROP-seq data processing
- ○ Differential gene expression analysis with mouse CROP-seq data
- ○ Cellular expression of mouse CROP-seq DEGs
- ○ Biological pathway analysis of mouse CROP-seq DEGs
- ○ Expression of the human orthologs of mouse CROP-seq DEGs in postmortem brains of individuals with schizophrenia
- ○ PPI network of mouse CROP-seq DEGs

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.12.022.

REFERENCES

1. GHDx, IHME (2019). Global Burden of Disease Study 2019 (GBD 2019) Data Resources. https://ghdx.healthdata.org/gbd-2019.

2. King, B.H. (2016). Psychiatric comorbidities in neurodevelopmental disorders. Curr. Opin. Neurol. 29, 113–117. https://doi.org/10.1097/WCO.0000000000000299.

3. Brainstorm Consortium, Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., and Gormley, P. (2018). Analysis of shared heritability in common disorders of the brain. Science 360, eaap8757. https://doi.org/10.1126/science.aap8757.

4. 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. https://doi.org/10.1176/appi.ajp.2017.17030283.

5. Cross-Disorder Group of the Psychiatric Genomics Consortium (2019). Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. Cell 179, 1469–1482.e11. https://doi.org/10.1016/j.cell.2019.11.020.

6. Hindley, G., Frei, O., Shadrin, A.A., Cheng, W., O'Connell, K.S., Icick, R., Parker, N., Bahrami, S., Karadag, N., Roelfs, D., et al. (2022). Charting the landscape of genetic overlap between mental disorders and related traits beyond genetic correlation. Am. J. Psychiatry 179, 833–843. https://doi.org/10.1176/appi.ajp.21101051.

7. McAfee, J.C., Bell, J.L., Krupa, O., Matoba, N., Stein, J.L., and Won, H. (2022). Focus on your locus with a massively parallel reporter assay. J. Neurodev. Disord. 14, 50. https://doi.org/10.1186/s11689-022-09461-x.

8. Sey, N.Y.A., Hu, B., Mah, W., Fauni, H., McAfee, J.C., Rajarajan, P., Brennand, K.J., Akbarian, S., and Won, H. (2020). A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. Nat. Neurosci. 23, 583–593. https://doi.org/10.1038/s41593-020-0603-0.

9. Spiess, K., and Won, H. (2020). Regulatory landscape in brain development and disease. Curr. Opin. Genet. Dev. 65, 53–60. https://doi.org/10.1016/j.gde.2020.05.007.

10. Schork, A.J., Won, H., Appadurai, V., Nudel, R., Gandal, M., Delaneau, O., Revsbech Christiansen, M., Hougaard, D.M., Bækved-Hansen, M., Bybjerg-Grauholm, J., et al. (2019). A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. Nat. Neurosci. 22, 353–361. https://doi.org/10.1038/s41593-018-0320-0.

11. Ziffra, R.S., Kim, C.N., Ross, J.M., Wilfert, A., Turner, T.N., Haeussler, M., Casella, A.M., Przytycki, P.F., Keough, K.C., Shin, D., et al. (2021). Single-cell epigenomics reveals mechanisms of human cortical development. Nature 598, 205–213. https://doi.org/10.1038/s41586-021-03209-8.

12. Khetan, S., Kales, S., Kursawe, R., Jillette, A., Ulirsch, J.C., Reilly, S.K., Ucar, D., Tewhey, R., and Stitzel, M.L. (2021). Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive β cell transcriptional activation. Nat. Commun. 12, 5242. https://doi.org/10.1038/s41467-021-25514-6.

13. Oliveros, W., Delfosse, K., Lato, D.F., Kiriakopulos, K., Mokhtaridoost, M., Said, A., McMurray, B.J., Browning, J.W.L., Mattioli, K., Meng, G., et al. (2023). Systematic characterization of regulatory variants of blood pressure genes. Cell Genomics 3, 100330. https://doi.org/10.1016/j.xgen.2023.100330.

14. Liang, L., Cao, C., Ji, L., Cai, Z., Wang, D., Ye, R., Chen, J., Yu, X., Zhou, J., Bai, Z., et al. (2023). Complementary Alu sequences mediate enhancer-promoter selectivity. Nature 619, 868–875. https://doi.org/10.1038/s41586-023-06323-x.

15. Wisniewska, M.B., Misztal, K., Michowski, W., Szczot, M., Purta, E., Lesniak, W., Klejman, M.E., Dabrowski, M., Filipkowski, R.K., Nagalski, A., et al. (2010). LEF1/beta-catenin complex regulates transcription of the Cav3.1 calcium channel gene (Cacna1g) in thalamic neurons of the adult brain. J. Neurosci. 30, 4957–4969. https://doi.org/10.1523/JNEUROSCI.1425-09.2010.

16. Duclot, F., and Kabbaj, M. (2017). The role of early growth response 1 (EGR1) in brain plasticity and neuropsychiatric disorders. Front. Behav. Neurosci. 11, 35. https://doi.org/10.3389/fnbeh.2017.00035.

17. McAfee, J.C., Lee, S., Lee, J., Bell, J.L., Krupa, O., Davis, J., Insigne, K., Bond, M.L., Zhao, N., Boyle, A.P., et al. (2023). Systematic investigation of allelic regulatory activity of schizophrenia-associated common variants. Cell Genomics 3, 100404. https://doi.org/10.1016/j.xgen.2023.100404.

18. Fullard, J.F., Giambartolomei, C., Hauberg, M.E., Xu, K., Voloudakis, G., Shao, Z., Bare, C., Dudley, J.T., Mattheisen, M., Robakis, N.K., et al. (2017). Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. Hum. Mol. Genet. 26, 1942–1951. https://doi.org/10.1093/hmg/ddx103.

19. Lalonde, S., Codina-Fauteux, V.-A., de Bellefon, S.M., Leblanc, F., Beaudoin, M., Simon, M.-M., Dali, R., Kwan, T., Lo, K.S., Pastinen, T., et al. (2019). Integrative analysis of vascular endothelial cell genomic features identifies AIDA as a coronary artery disease candidate gene. Genome Biol. 20, 133. https://doi.org/10.1186/s13059-019-1749-5.

20. Bysani, M., Agren, R., Davegårdh, C., Volkov, P., Rönn, T., Unneberg, P., Bacos, K., and Ling, C. (2019). ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes. Sci. Rep. 9, 7785. https://doi.org/10.1038/s41598-019-44076-8.

21. Ibarra, I.L., Hollmann, N.M., Klaus, B., Augsten, S., Velten, B., Hennig, J., and Zaugg, J.B. (2020). Mechanistic insights into transcription factor co-operativity and its impact on protein-phenotype interactions. Nat. Commun. *11*, 124. https://doi.org/10.1038/s41467-019-13888-7.

22. Maioli, S., Leander, K., Nilsson, P., and Nalvarte, I. (2021). Estrogen receptors and the aging brain. Essays Biochem. *65*, 913–925. https://doi.org/10.1042/EBC20200162.

23. Mey, J., and McCaffery, P. (2004). Retinoic acid signaling in the nervous system of adult vertebrates. Neuroscientist *10*, 409–421. https://doi.org/10.1177/1073858404263520.

24. Benner, C., Spencer, C.C.A., Havulinna, A.S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics *32*, 1493–1501. https://doi.org/10.1093/bioinformatics/btw018.

25. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the "Sum of Single Effects" model. PLoS Genet. *18*, e1010299. https://doi.org/10.1371/journal.pgen.1010299.

26. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497–508. https://doi.org/10.1534/genetics.114.167908.

27. de Klein, N., Tsai, E.A., Vochteloo, M., Baird, D., Huang, Y., Chen, C.-Y., van Dam, S., Oelen, R., Deelen, P., Bakker, O.B., et al. (2023). Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. Nat. Genet. *55*, 377–388. https://doi.org/10.1038/s41588-023-01300-6.

28. Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V.A., Selvaraj, S., Nutma, E., Marzin, M., et al. (2022). Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. Nat. Neurosci. *25*, 1104–1112. https://doi.org/10.1038/s41593-022-01128-z.

29. 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. https://doi.org/10.1126/science.aat8464.

30. Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. *36*, 70–80. https://doi.org/10.1038/nbt.4038.

31. BRAIN Initiative Cell Census Network (BICCN) (2021). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature *598*, 86–102. https://doi.org/10.1038/s41586-021-03950-0.

32. Emani, P.S., Liu, J.J., Clarke, D., Jensen, M., Warrell, J., Gupta, C., Meng, R., Lee, C.Y., Xu, S., Dursun, C., et al. (2024). Single-cell genomics and regulatory networks for 388 human brains. Science *384*, eadi5199. https://doi.org/10.1126/science.adi5199.

33. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008–1021. https://doi.org/10.1016/j.cell.2013.10.031.

34. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215. https://doi.org/10.1038/nature13772.

35. Network and Pathway Analysis Subgroup of Psychiatric Genomics Consortium (2015). Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. Nat. Neurosci. *18*, 199–209. https://doi.org/10.1038/nn.3922.

36. 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. https://doi.org/10.1038/s41588-022-01104-0.

37. 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. https://doi.org/10.1038/s41586-020-2832-5.

38. Mah, W., and Won, H. (2020). The three-dimensional landscape of the genome in human brain tissue unveils regulatory mechanisms leading to schizophrenia risk. Schizophr. Res. *217*, 17–25. https://doi.org/10.1016/j.schres.2019.03.007.

39. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301. https://doi.org/10.1038/nmeth.4177.

40. Tian, R., Gachechiladze, M.A., Ludwig, C.H., Laurie, M.T., Hong, J.Y., Nathaniel, D., Prabhu, A.V., Fernandopulle, M.S., Patel, R., Abshari, M., et al. (2019). CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron *104*, 239–255.e12. https://doi.org/10.1016/j.neuron.2019.07.014.

41. Fregeau, B., Kim, B.J., Hernández-García, A., Jordan, V.K., Cho, M.T., Schnur, R.E., Monaghan, K.G., Juusola, J., Rosenfeld, J.A., Bhoj, E., et al. (2016). De Novo Mutations of RERE Cause a Genetic Syndrome with Features that Overlap Those Associated with Proximal 1p36 Deletions. Am. J. Hum. Genet. *98*, 963–970. https://doi.org/10.1016/j.ajhg.2016.03.002.

42. Vosberg, D.E., Leyton, M., and Flores, C. (2020). The Netrin-1/DCC guidance system: dopamine pathway maturation and psychiatric disorders emerging in adolescence. Mol. Psychiatry *25*, 297–307. https://doi.org/10.1038/s41380-019-0561-7.

43. Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–489. https://doi.org/10.1038/nature10523.

44. Murphy, K.E., Meng, F.W., Makowski, C.E., and Murphy, P.J. (2020). Genome-wide chromatin accessibility is restricted by ANP32E. Nat. Commun. *11*, 5063. https://doi.org/10.1038/s41467-020-18821-x.

45. Liao, T., Wang, Y.-J., Hu, J.-Q., Wang, Y., Han, L.-T., Ma, B., Shi, R.-L., Qu, N., Wei, W.-J., Guan, Q., et al. (2018). Histone methyltransferase KMT5A gene modulates oncogenesis and lipid metabolism of papillary thyroid cancer in vitro. Oncol. Rep. *39*, 2185–2192. https://doi.org/10.3892/or.2018.6295.

46. Lanouette, S., Haddad, J., Zhang, P., and Couture, J.F. (2016). Impacts of Histone Lysine Methylation on Chromatin. In Chromatin signaling and diseases (Elsevier), pp. 25–53. https://doi.org/10.1016/B978-0-12-802389-1.00002-2.

47. Platt, R.J., Chen, S., Zhou, Y., Yim, M.J., Swiech, L., Kempton, H.R., Dahlman, J.E., Parnas, O., Eisenhaure, T.M., Jovanovic, M., et al. (2014). CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell *159*, 440–455. https://doi.org/10.1016/j.cell.2014.09.014.

48. Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323. https://doi.org/10.1126/science.aap8809.

49. Ruzicka, W.B., Mohammadi, S., Fullard, J.F., Davila-Velderrain, J., Subburaju, S., Tso, D.R., ihan, M., Jiang, S., Lee, H.-C., Bendl, J., et al. (2024). Single-cell multi-cohort dissection of the schizophrenia transcriptome. Science *384*, eadg5136. https://doi.org/10.1126/science.adg5136.

50. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

51. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., and Huber, W. (2005). BioMart and Bioconductor: a powerful link between

biological databases and microarray data analysis. Bioinformatics *21*, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525.

52. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

53. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

54. Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., UK10K Consortium, Dunham, I., and Birney, E. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. *51*, 343–353. https://doi.org/10.1038/s41588-018-0322-6.

55. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., and Peterson, H. (2020). gprofiler2—an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res. *9*. https://doi.org/10.12688/f1000research.24956.2.

56. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

57. Coetzee, S.G., Coetzee, G.A., and Hazelett, D.J. (2015). motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics *31*, 3847–3849. https://doi.org/10.1093/bioinformatics/btv470.

58. Myint, L., Avramopoulos, D.G., Goff, L.A., and Hansen, K.D. (2019). Linear models enable powerful differential activity analysis in massively parallel reporter assays. BMC Genomics *20*, 209. https://doi.org/10.1186/s12864-019-5556-x.

59. Kolde, R. (2024). pheatmap: Pretty Heatmaps. The R Foundation. https://doi.org/10.32614/cran.package.pheatmap.

60. 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. https://doi.org/10.1186/s13742-015-0047-8.

61. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

62. Lawrence, M., Gentleman, R., and Carey, V. (2009). rtracklayer: an R package for interfacing with genome browsers. Bioinformatics *25*, 1841–1842. https://doi.org/10.1093/bioinformatics/btp328.

63. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat. Biotechnol. *42*, 293–304. https://doi.org/10.1038/s41587-023-01767-y.

64. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. https://doi.org/10.1038/s41590-018-0276-y.

65. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.

66. Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics *33*, 2938–2940. https://doi.org/10.1093/bioinformatics/btx364.

67. Davis, J.E., Insigne, K.D., Jones, E.M., Hastings, Q.A., Boldridge, W.C., and Kosuri, S. (2020). Dissection of c-AMP Response Element Architecture by Using Genomic and Episomal Massively Parallel Reporter Assays. Cell Syst. *11*, 75–85.e7. https://doi.org/10.1016/j.cels.2020.05.011.

68. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv. https://doi.org/10.48550/arxiv.1303.3997.

69. 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. https://doi.org/10.1093/bioinformatics/btt656.

70. Kojima, K.K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. Mobile DNA *9*, 2. https://doi.org/10.1186/s13100-017-0107-y.

71. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., and McVean, G.A. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

72. Storey, J.D. (2011). False Discovery Rate. In International encyclopedia of statistical science, M. Lovric, ed. (Springer Berlin Heidelberg), pp. 504–508. https://doi.org/10.1007/978-3-642-04898-2_248.

73. Liu, S., Won, H., Clarke, D., Matoba, N., Khullar, S., Mu, Y., Wang, D., and Gerstein, M. (2022). Illuminating links between cis-regulators and trans-acting variants in the human prefrontal cortex. Genome Med. *14*, 133. https://doi.org/10.1186/s13073-022-01133-8.

74. Fulco, C.P., Nasser, J., Jones, T.R., Munson, G., Bergman, D.T., Subramanian, V., Grossman, S.R., Anyoha, R., Doughty, B.R., Patwardhan, T.A., et al. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nat. Genet. *51*, 1664–1669. https://doi.org/10.1038/s41588-019-0538-0.

75. Hu, B., Won, H., Mah, W., Park, R.B., Kassim, B., Spiess, K., Kozlenkov, A., Crowley, C.A., and Pochareddy, S.; PsychENCODE Consortium (2021). Neuronal and glial 3D chromatin architecture informs the cellular etiology of brain disorders. Nat. Commun. *12*, 3968. https://doi.org/10.1038/s41467-021-24243-0.

76. 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. https://doi.org/10.1038/nature19847.

77. Aygün, N., Elwell, A.L., Liang, D., Lafferty, M.J., Cheek, K.E., Courtney, K.P., Mory, J., Hadden-Ford, E., Krupa, O., de la Torre-Ubieta, L., et al. (2021). Brain-trait-associated variants impact cell-type-specific gene regulation during neurogenesis. Am. J. Hum. Genet. *108*, 1647–1668. https://doi.org/10.1016/j.ajhg.2021.07.011.

78. Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222–3241.e26. https://doi.org/10.1016/j.cell.2021.04.021.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
AlexaFluor488 anti-mouse	Thermo Fisher	Cat#A-28175; RRID: AB_2536161
AlexaFluor647 anti-rabbit	Thermo Fisher	Cat#A-21245; RRID: AB_2535813
Mouse MAP2 primary antibody	Millipore Sigma	Cat#M1406; RRID: AB_477171
Mouse NF-H primary antibody	Thermo Fisher	Cat#MA1-2012; RRID: AB_347005
Mouse TUJ1 primary antibody	Millipore Sigma	Cat#T8578; RRID: AB_1841228
Rabbit PSD95 primary antibody	Millipore Sigma	Cat#P246; RRID: AB_260911
Rabbit SATB2 primary antibody	abcam	Cat#ab34735; RRID: AB_2301417
Bacterial and virus strains		
CROPseq-tRFP-guideLib-invitro	UNC Lenti-shRNA Core Facility	N/A
CROPseq-tRFP-guideLib-invivo	UNC Lenti-shRNA Core Facility	N/A
Endura electrocompetent cells	Lucigen	Cat#60242-1
Chemicals, peptides, and recombinant proteins		
AMPure XP Beads	Beckman Coulter	Cat#A63881
Avantor® Seradigm, Premium Grade Fetal Bovine Serum (FBS)	Avantor	Cat#97068-085
B-27™ Plus Supplement (50X)	Thermo Fisher	Cat#A3582801
B-27™ Supplement (50X), minus vitamin A	Thermo Fisher	Cat#12587010
BIT 9500	STEMCELL	Cat#9500
Brain-Derived Neurotrophic Factor (BDNF)	PeproTech	Cat#450-02
BSA	Millipore Sigma	Cat#A9576
DAPI	Millipore Sigma	Cat#D9542
DMEM	Thermo Fisher	Cat#10313-039
DMEM/F12, HEPES	Thermo Fisher	Cat#11330-032
Doxycycline hydrochloride	Sigma-Aldrich	Cat#D3447-500MG
DPBS (without Ca2+ or Mg2+) 1X	Corning	Cat#20-031-CV
Epidermal growth factor (EGF)	PeproTech	Cat#AF-100-15
Fetal Bovine Serum (FBS)	Corning	Cat#35-015-CV
Fibroblast growth factors (FGF)	PeproTech	Cat#AF-100-15
Formaldehyde solution	Millipore Sigma	Cat#F1635
GlutaMax 100X	Thermo Fisher	Cat#35050-061
Hank's Balanced Salt Solution (HBSS) 1X	Corning	Cat#21-022-CV
Heparin	Sigma-Aldrich	Cat#H3393-100KU
Hibernate A Low-Fluorescence	BrainBits	Cat#HALF500
Knockout DMEM	Thermo Fisher	Cat#10829-018
Knockout DMEM/F12	Thermo Fisher	Cat#12660-012
KpnI-HF	New England Biolabs (NEB)	Cat#R3142S
Leukemia inhibitory factor (LIF)	PeproTech	Cat#300-05
Lipofectamine STEM	Invitrogen	Cat#STEM00015
Magnesium Chloride 1M	Thermo Fisher	Cat#AM9530G
Matrigel Matrix	Corning	Cat#356231
MluI-HF	New England Biolabs (NEB)	Cat#R3198S
Mouse Laminin	Thermo Fisher	Cat#23017-015
N2 Supplement 100X	Thermo Fisher	Cat#17502-048
NEAA (MEM Non-Essential Amino Acids)	Thermo Fisher	Cat#11140-050

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
NEBNext 2X Q5 Hifi HS Master Mix	New England Biolabs (NEB)	Cat#M0453S
Neurobasal-A media	Thermo Fisher	Cat#10888022
Neurotrophin-3 (NT-3)	PeproTech	Cat#450-03
Nucleic acid transfection enhancer (NATE)	InvivoGen	Cat#lyec-nate
Papain Dissociation System	Worthington	Cat#LK003153
Platelet-derived growth factor (PDGF)	PeproTech	Cat#100-00AB
Poly-L-Ornithine	Sigma-Aldrich	Cat#P3655-100MG
Primocin	Invitrogen	Cat#ant-pm-2
ROCK inhibitor (Y-27632 2HCl)	Selleck Chemicals	Cat#S1049
ROCK inhibitor (Y-27632 2HCl)	Tocris	Cat#125410
SparQ beads	VWR	Cat#76302-834
SpeI-HF	New England Biolabs (NEB)	Cat#R3133S
StemFlex™ Medium	Thermo Fisher	Cat#A3349401
StemPro Accutase Cell Dissociation Reagent	Thermo Fisher	Cat#A11105-01
SuperScript IV Reverse Transcriptase	Invitrogen	Cat#18090050
T7 DNA ligase	New England Biolabs (NEB)	Cat#M0318S
Trimethoprim (TMP)	Sigma-Aldrich	Cat#92131-1G
Triton	Millipore Sigma	Cat#T8787
Tween20	Milipore Sigma	Cat#P2287
XbaI	New England Biolabs (NEB)	Cat#R0145S
Critical commercial assays		
DNA Clean and Concentrator Kit-25	Zymo Research	Cat#D4033
Qiagen Plasmid Maxi Kit-25	Qiagen	Cat#12162
Qiagen Plasmid Mini Kit-25	Qiagen	Cat#12123
Quick-DNA/RNA Miniprep Plus Kit	Zymo Research	Cat#D7003
Zymo DNA Clean and Concentrator Kit-5	Zymo Research	Cat#D4033
Zymo Gel DNA Recovery Kit	Zymo Research	Cat#D4008
ZymoPURE II Plasmid Maxiprep Kit	Zymo Research	Cat#D4203
Deposited data		
Code for analysis	This Paper	https://github.com/thewonlab/crossdisorder-MPRA
CROP-seq in vitro (RT029 hiPSC-derived Neurons)	This Paper	GEO: GSE276947
CROP-seq in vivo	This Paper	GEO: GSE282731
Raw MPRA data	This Paper	GEO: GSE244011
Experimental models: Cell lines		
Human neural progenitors (HNPs)	Dr. Jason Stein's lab, UNC-CH	Donor #88
Human: i3N RT029 iPSC line (inducible CRISPRi, inducible NGN2)	Dr. Martin Kampmann, UCSF[40]	RT029/SF2019-227
Experimental models: Organisms/strains		
Mouse: Rosa26-Cas9 knockin on B6J	Jackson	RRID:IMSR_JAX:026179
Oligonucleotides		
Sequences in Table S6	IDT	N/A
Recombinant DNA		
Plasmid: CROPseq-tRFP	This Paper	Addgene #230938
Plasmid: Donor_eGP2AP_RC	Addgene	Plasmid#133784
Plasmid: pMPRAdonor2 plasmid	Addgene	Plasmid#49353
Plasmid: pUC19	Thermo Fisher	Cat#SD0061

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
10x Genomics Cell Ranger v6.1.2	Zheng et al.[50]	https://www.10xgenomics.com/support/software/cell-ranger/latest
10x Genomics Cell Ranger v7.0.0	Zheng et al.[50]	https://www.10xgenomics.com/support/software/cell-ranger/latest
BioMart (version 2.58.0)	Durinck et al.[51]	https://bioconductor.org/packages/release/bioc/html/biomaRt.html
BWA (version 0.7.17-r1188)	Li et al.[52]	https://bio-bwa.sourceforge.net/
CAVIAR (version 2.2)	Hormozdiari et al.[26]	https://github.com/fhormoz/caviar
coloc (version 5.1.0.1)	Giambartolomei et al.[53]	https://github.com/chr1swallace/coloc
FINEMAP (version 1.4.2)	Benner et al.[24]	http://www.christianbenner.com/
garfield (version 2.0)	Iotchkova et al.[54]	https://www.bioconductor.org/packages/release/bioc/html/garfield.html
gprofiler2 (version 0.2.3)	Kolberg et al.[55]	https://biit.cs.ut.ee/gprofiler/page/r
HOMER (version 4.10.4)	Heinz et al.[56]	http://homer.ucsd.edu/homer/motif/
motifbreakR (version 2.10.2)	Coetzee et al.[57]	https://bioconductor.org/packages/release/bioc/html/motifbreakR.html
mpra (version 1.24.0)	Myint et al.[58]	https://www.bioconductor.org/packages/release/bioc/html/mpra.html
pheatmap (version 1.0.12)	Kolde[59]	https://cran.r-project.org/web/packages/pheatmap/index.html
PLINK2 (version 1.90b3.45)	Chang et al.[60]	https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533
RepeatMasker (version Open-3.0)	Tarailo-Graovac et al.[61]	https://github.com/Dfam-consortium/RepeatMasker
rtracklayer (version 1.54.0)	Lawrence et al.[62]	https://bioconductor.org/packages/release/bioc/html/rtracklayer.html
Seurat (version 5.0.2)	Hao et al.[63]	https://satijalab.org/seurat/
singleR (version 2.2.0)	Aran et al.[64]	https://github.com/SingleR-inc/SingleR
STRING (version 12.0)	Szklarczyk et al.[65]	https://string-db.org/
susieR (version 0.12.35)	Zou et al.[25]	https://github.com/stephenslab/susieR
UpsetR (version 1.4.0)	Conway et al.[66]	https://github.com/hms-dbmi/UpSetR

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Protocols for HNP acquisition, generation, and cell culture were previously described by McAfee et al.[17] Donor #88 (genetically male, HNPs obtained at gestation week 14) was used for all experiments. Briefly, 800,000 cells were plated per 6-well plate pre-coated with Poly-L-Ornithine (10 μg/ml; Sigma-Aldrich, cat. no. P3655-100MG) and laminin (5 μg/ml; Gibco, cat. no. 23017015). Cells were cultured in Neurobasal-A media (Thermo Fisher, cat. no. 10888022) supplemented with primocin (100 μg/ml; Invitrogen, cat. no. ant-pm-2), BIT 9500 (10%; STEMCELL, cat. no. 09500), Glutamax 100X (1%; Fisher Scientific, cat. no. 5112367), heparin (1 μg/ml; Sigma-Aldrich, cat. no. H3393-100KU), and growth factors: FGF (20 μg/ml; PeproTech, cat. no. AF-100-15), LIF (2 ng/ml; PeproTech, cat. no. 300-05), EGF (20 μg/ml; PeproTech, cat. no. AF-100-15), and PDGF (20 ng/ml; PeproTech, cat. no. 100-00AB). Cells were maintained in a 5% CO_2 incubator at 37°C.

WTC11 i³Neuron dCas9-KRAB hiPSCs (genetically male iPSC line with inducible *NGN2* and CRISPRi, WTC11 background, RT029/SF2019-227, obtained from the Kampmann Lab) were differentiated into glutamatergic neurons following the iNeuron pre-differentiation and differentiation protocol from the Kampmann lab.[40] Briefly, hiPSCs were cultured for about a week in Stemflex media (Thermo Fisher, cat. no. A3349401). Cells were dissociated using accutase (Thermo Fisher, cat. no. A11105-01) and seeded at 7.5×10^5 cells/well in 6-well plates in pre-differentiation media on day –3 for pre-differentiation into neural progenitor cells (NPCs). Half media changes with N2 pre-differentiation media were performed daily until day 0. On day 0, cells were dissociated once again using accutase, and re-plated at 7×10^6 cells/well in Poly-D-Lysine (PDL; Thermo Fisher, cat. no. A3890401)-coated 10-cm dishes with N2/B27 differentiation media for differentiation into neurons. The pre-differentiation media consisted of Knockout DMEM/F12

(Thermo Fisher, cat. no. 12660-012), NEAA 100X (Thermo Fisher, cat. no. 11140-050), and N2 Supplement S100X (Thermo Fisher, cat. no. 17502-048). Differentiation media consisted of DMEM/F12 HEPES (Thermo Fisher, cat. no. 11330-032), Neurobasal-A media, NEAA 100X, GlutaMax 100X (Thermo Fisher, cat. no. 35050-061), N2 Supplement 100X, and B27 minus vitamin A 50X (Thermo Fisher, cat. no. 12587010). Trimethoprim (Sigma-Aldrich, cat. no. 92131-1G), NT-3 (PeproTech, cat. no. 450-03), BDNF (PeproTech cat. no. 450-02), and mouse laminin (Thermo Fisher, cat. no. 23017-015) were added during plating and media changes. ROCK inhibitor (Tocris, cat. no. 125410) was added only during plating on day –3 and day 0. Finally, Doxycycline (Sigma-Aldrich, cat. no. D3447-500MG) was added during plating on day –3, during media changes on day –2 and day –1, and during plating on day 0. Cells were grown in a 5% CO_2 incubator at 37°C.

All mouse work was performed under the guidelines of UNC-Chapel Hill Institutional Animal Care and Use Committee (IACUC), complying with the IACUC-approved protocol 23-008.0. Cas9 transgenic mice[47] (JAX cat. no. RRID:IMSR_JAX:026179) were used for *in vivo* CROP-seq experiments. The mice were maintained in a C57BL/6J background. All mice used for the experiment were heterozygous for the Cas9 transgene, obtained by crossing homozygous males to wild-type C57BL/6J females. After 3 days of mating, dams were group-housed until 3 to 5 days before giving birth. Dams were single-housed by the time pups were born. Pups stayed with their mother until the single-cell prep experiment (P13-P15). A total of 15 pups (10 males, 5 females) from 8 different litters were used for the dataset.

METHOD DETAILS

Cross-disorder variant selection
Variants that overlap 136 GWS loci with $p < 1 \times 10^{-5}$ were selected from the cross-disorder GWAS.[5] Variant sequences that harbor restriction enzyme sites used for downstream molecular cloning (MluI, SpeI, KpnI, XbaI) were removed, leaving 17,841 variants to be tested. Alleles were positioned in the middle of a 150-bp region of their surrounding genomic sequences. Tiled 150-bp sections of the CMV and EF1α promoters were included as positive controls (with a total of 6 sequences), and sequences of scrambled DNA with matching GC contents to the MPRA library were included as negative controls (with a total of 100 sequences). We used Agilent oligoarray chip technology to synthesize 150-bp variant-harboring sequences (sequences extracted from hg19) with flanking restriction enzyme sites for molecular cloning and primer sequences for amplification (total 200-bp in size).

MPRA library construction
The synthesized 200-bp oligo library was diluted to 1:20 and amplified using MPRA-chip-primer-F and MPRA-chip-primer-R with NEBNext 2X Q5 Hifi HS Mastermix (NEB, cat. no. M0453S). The resulting PCR reaction was cleaned up with Zymo DNA Clean and Concentrator Kit-5 (Zymo Research, cat. no. D4033). Random barcodes were added via PCR using MPRA-bc-primer-F and MPRA-bc-primer-R with NEBNext 2X Q5 Hifi HS Mastermix. The resulting PCR product was cleaned up with Zymo DNA Clean and Concentrator Kit-5. The MPRA vector, Donor_eGP2AP_RC (Addgene: ID#133784),[67] and the oligo library was digested with SpeI-HF (NEB, cat. no. R3133S) and MluI-HF (NEB, cat. no. R3198S). The digested oligo library was ligated into the Donor_eGP2AP_RC plasmid using T7 DNA ligase (NEB, cat. no. M0318S). The ligation mix was transformed into Endura Electrocompetent cells (Lucigen, cat. no. 60242-1) via electroporation. After 1 hour of recovery at 37°C, the bacteria was plated onto 10 cm LB plates with kanamycin and incubated overnight at 37°C. Colonies were scraped and cultured in 4L of LB with kanamycin at 30°C overnight. Plasmid DNA was purified using ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, cat. no. D4203).

The library and barcode region were amplified from the plasmid with BC_Map_P5_R and BC_Map_P7_F_Ind_# using NEBNext 2X Q5 Hifi HS Mastermix, which added sequencing adapters and indices for demultiplexing. This PCR product was size-selected and cleaned up using SparQ beads (VWR, cat. no. 76302-834), and sequenced for barcode mapping via 2x150-bp Novaseq SP with 20% PhiX and custom sequencing primers: BC_Map_R1seq, BC_Map_R2seq, and BC_Map_Index. Barcodes were assigned to each oligo sequence, as previously described[17] using the custom code: https://github.com/kiminsigne-ucla/bc_map. See Table S6 for primer sequences.

The minimal promoter (minP) and luciferase (Luc) sequence was digested from the pMPRAdonor2 plasmid (Addgene: ID#49353) using KpnI-HF (NEB, cat. no. R3142S) and XbaI (NEB, cat. no. R0145S). The desired band was extracted from a 1% agarose gel using Zymo Gel DNA Recovery Kit (Zymo Research, cat. no. D4008). The Donor_eGP2AP_RC plasmid containing the library and barcode was digested with KpnI-HF and XbaI and cleaned up with the Zymo DNA Clean and Concentrator Kit-5. The digested plasmid and the minP-Luc fragment were ligated using T7 DNA ligase. The ligase reaction was transformed into Endura Electrocompetent cells, plated on 10 cm LB plates with kanamycin, and grown at 37°C overnight. Colonies were scraped and cultured in 4L of LB with kanamycin at 30°C overnight. The plasmid was purified using the ZymoPURE II Plasmid Maxiprep Kit.

Introduction of MPRA libraries into HNPs
HNPs were plated and maintained as described in the experimental model and study participant details. Forty-eight hours after plating, cells were half-fed using the Neurobasal-A media supplemented with primocin, BIT 9500, Glutamax 100X, and heparin with twice the amount of growth factors and nucleic acid transfection enhancer (NATE) at 1:50 final dilution (InvivoGen, cat. no. lyec-nate). We transfected each well with 2.5 ug of MPRA plasmid library and 1 ug of pUC19 (Thermo Fisher, cat. no. SD0061) using Lipofectamine STEM (Invitrogen, cat. no. STEM00015) according to the manufacturer's instructions. Forty-eight hours after

transfection, cells were half-fed. Seventy-two hours after transfection, DNA and RNA were extracted from the cells by Quick-DNA/RNA Miniprep Plus Kit (Zymo Research, cat. no. D7003) using 600 uL of shield buffer and 600 uL of lysis buffer per well.

RNA and DNA library generation

Extracted RNA was reverse transcribed into cDNA using SuperScript IV Reverse Transcriptase (Invitrogen, cat. no. 18090050) with a primer targeting the region downstream of the barcodes (Lib_Hand_RT). The resulting cDNA was amplified with primers Lib_Hand and Lib_seq_Luc_R using NEBNext 2X Q5 Hifi HS Mastermix. The PCR product was cleaned up with Zymo DNA Clean and Concentrator Kit-25 (Zymo Research, cat. no. D4033). In parallel, 4 uL of DNA from each well was amplified by 4 PCR reactions using primers (Lib_Hand_RT and Lib_Seq_Luc_R) and NEBNext 2X Q5 Hifi HS Mastermix. The PCR product was similarly cleaned up with Zymo Clean and Concentrator Kit-25. We then added P5/P7 adapter sequences and demultiplexing indices to both the cDNA and DNA libraries by PCR using primers (P5_seq_Luc_F and P7_Ind_#_Han) and NEBNext 2X Q5 Hifi HS Mastermix. The PCR product was cleaned up and size-selected using AMPure XP Beads (Beckman Coulter, cat. no. A63881) with 0.7X and 0.9X ratios to select for a DNA fragment at 273-bp. The libraries were subsequently sequenced via NovaSeq SP using custom cycles of 35 x 8 x 0 with 20% PhiX and custom sequencing primers. See Table S6 for primer sequences.

CROP-seq in hiPSC-derived neurons

The CROP-seq library was generated using a modified version of the original CROP-seq vector[39] with the puromycin sequence replaced by tag red fluorescent protein (tRFP) (CROPseq-tRFP, Addgene: ID#230938). The gRNAs targeting three emVars (rs4513167 and rs4614799 in the *DCC* locus; rs301804 in the *RERE* locus), along with two additional MPRA-allelic/inactive variant targets (rs8089270 and rs6508210 in the *DCC* locus) were inserted into the CROP-seq vector, as previously described.[39] Four gRNAs were designed per variant to account for variability in individual gRNA knockdown efficiency (see Table S6 for gRNA sequences). The CROP-seq library was then packaged into lentivirus (CROPseq-tRFP-guideLib-in vitro; Vector copy number [VCN] titer: 2.07x10[10] transducing units [TU]/mL) by the UNC Lenti-shRNA Core Facility.

CROP-seq in hiPSC-derived neurons

WTC11 i³Neuron dCas9-KRAB hiPSCs were grown and differentiated as described in the experimental model and study participant details section. CROP-seq lentivirus was added to cells on day 0 at a multiplicity of infection (MOI) of 4. The lentivirus was removed via media change after 24 hours of transduction. Fluorescence of RFP was monitored to ensure successful transduction of cells. When neurons reached day 11 of differentiation, they were collected from wells through single-cell dissociation using the Worthington Papain Dissociation System (Worthington, cat. no. LK003153). Briefly, on the day of collection, papain (Worthington, cat. no. LK003153) was resuspended in 5 mL 1X HBSS (Thermo, cat. no. 2277417) to a final concentration of 20 U/mL. A resuspension solution of DMEM (Thermo Fisher, cat. no. 10313039) with 10% FBS was also prepared, and both solutions were pre-warmed at 37°C for at least 10 minutes. Conditioned media from the neuron cultures was collected separately from lentivirus-treated and untreated cells. To enhance cell survival and prevent aggregation during FACS preparation, 10 μM Y-27632 dihydrochloride ROCK inhibitor and 2% FBS (Corning, cat. no. 35-015-CV) were added to each tube of conditioned media. Immediately before use, DNase (Worthington, cat. no. LK003153) at 5 μg/mL and Magnesium Chloride (Thermo Fisher, cat. no. AM9530G) at 5 mM were added to the papain solution. The culture media was aspirated, cells were washed with DPBS (Thermo Fisher, cat. no. 14190144), and the papain dissociation solution was added at 500 uL per well. The plate was then placed in the 37°C incubator for 10 minutes and after incubation, gently agitated to ensure neuron sheet dissociation. The papain dissociation solution was then quenched in 5 volumes of DMEM and 10% FBS solution, and the neuron sheet was then gently triturated with a P1000 pipette. Lentivirus-treated and untreated cells were transferred separately to 15 mL conical tubes and centrifuged at 200g for 10 minutes. The supernatant was carefully removed, and cells were resuspended by slow vortexing in the conditioned media solution containing ROCK inhibitor and 2% FBS in preparation for FACS.

Live, RFP-expressing cells were isolated via FACS by the UNC Advanced Analytics (AA) Core. After FACS, RFP+ cells were immediately processed by the UNC AA Core using the 10x Genomics 3' v3.1 Dual Index workflow for scRNA-seq library generation. To assign gRNAs to cells, gRNA reads were separately enriched from cDNA using custom-designed primers targeting the CROP-seq backbone sequence (GuideEnrich_1, GuideEnrich_2, GuideEnrich_3, and GuideEnrich_4; see Table S6 for primer sequences). The resulting libraries were sequenced with Illumina NovaseqX 10B, using 3 lanes for the scRNA-seq libraries (cycles R1 x i1 x i2 x R2 = 28 x 10 x 10 x 90, with 1% PhiX) and 1 lane for gRNA enrichment libraries (cycles R1 x i1 x i2 x R2 = 28 x 10 x 10 x 42, with 20% PhiX).

Immunofluorescence staining of hiPSC-derived neurons

For immunofluorescence imaging, 12-mm diameter coverslips (Electron microscopy sciences, cat. no. 72230-01) were placed in 24-well plates and coated with PDL. Cells were seeded at a density of 1x10[5] cells/well on day 0 of plating. On day 11 of differentiation, the media was removed and cells were washed once with DPBS. Freshly prepared 4% formaldehyde (Millipore Sigma, cat. no. F1635) was added to wells (250 uL/well). The cells were then incubated at room temperature for 15 minutes. The formaldehyde was then removed and 250 uL DPBS were added 3 times to wash fixed cells. Afterwards, cells were permeabilized using 500 uL/well of 0.3% Triton (Millipore Sigma, cat. no. T8787)/PBS and incubated 15 minutes at room temperature. The Triton/PBS was

then aspirated from coverslips, followed by washing 3 times with DPBS. Blocking was performed by adding 250 uL/well of 5% BSA (Millipore Sigma, cat. no. A9576)/PBS and incubating at room temperature for 1 hour. All primary antibodies were diluted in 1% BSA/ PBS. They were mouse TUJ1 (Millipore Sigma, cat. no. T8578) diluted at 1:1000, rabbit SATB2 (abcam, cat. no. ab34735) diluted at 1:500, mouse MAP2 (Millipore Sigma, cat. no. M1406) diluted at 1:1000, mouse NF-H (Thermo Fisher, cat. no. MA1-2012) diluted at 1:1000, and rabbit PSD95 (Millipore Sigma, cat. no. P246) diluted at 1:500. Primary antibodies were added to the coverslips at 100 uL/ well and incubated overnight at $4°C$. Following incubation, the antibodies were removed, and the coverslips were washed 3 times with 0.1% Tween20 (Milipore Sigma, cat. no. P2287)/PBS (500 uL/well) for 5 minutes at room temperature per wash. All secondary antibodies were diluted 1:1000 in 1% BSA/PBS. They were AlexaFluor488 anti-mouse (Thermo Fisher, cat. no. A-28175), which was used for MAP2, TUJ1, and NF-H primary antibodies, and AlexaFluor647 anti-rabbit (Thermo Fisher, cat. no. A-21245), which was used for SATB2 and PSD95 primary antibodies. Secondary antibodies were added to the coverslips at 250 uL/well and incubated for 30 minutes at room temperature. After incubation, secondary antibodies were removed, and the coverslips were washed 3 times with 100 uL of 0.1% Tween20/PBS for 5 minutes at room temperature per wash. Finally, DAPI (Millipore Sigma, cat. no. D9542) was diluted 1:1000 in PBS and added to wells at 250 uL/well, incubating for 15 minutes at room temperature. The DAPI was removed and coverslips were washed 3 times with 250 uL of PBS for 5 minutes at room temperature per wash.

CROP-seq in mouse brain

gRNAs targeting the coding sequence of *Anp32e* (Anp32e_gRNA1 and Anp32e_gRNA2) and *Kmt5a* (Kmt5a_gRNA1 and Kmt5a_gRNA2) were inserted into CROPseq-tRFP using the previously described protocol.[39] gRNA sequences are described in Table S6.

The CROP-seq library was packaged into lentivirus (CROPseq-tRFP-guideLib-invivo, titer: 2.25 x 10^9 – 2.38 x 10^{10} TU/mL) by the UNC Lenti-shRNA Core Facility. The mice were injected with the lentivirus bilaterally into the lateral ventricles of the brain (2–3 uL of virus per mouse) at postnatal day 0. The virus was mixed with Fast Green FCF dye (Sigma, cat. no. F7252-5G, 0.5 mg/mL) for visual tracking of the injection then loaded into a Hamilton syringe (Hamilton, cat. no. 7634-01) with a 32 gauge needle (Hamilton, cat. no. 7803-04, point style 2, 0.4 inch length) attached to a custom stereotaxic instrument (Kopf, cat. no. 176-61-SB-A, 1449-A, 1772-A). Pups were anesthetized on ice for 2 minutes before being injected with the virus-dye mixture. They are then warmed up via a heating pad before being returned to the mother.

When the mice reached postnatal day 13-15, the neocortex was dissected for live single-cell dissociation using the Worthington Papain Dissociation System. Upon sacrifice, pups' toes were cut for genotyping confirmation of the Cas9 transgene. Genotyping PCR primers and protocols were from Jackson standard protocol (JAX, cat. no. RRID:IMSR_ JAX:026179). Briefly, minced neocortical tissue was incubated in papain/DNase solution with constant agitation for 90 minutes at $37°C$ in a 5% CO_2 incubator. After chemical digestion, the tissue was gently triturated using a serological pipet. Dissociated cells were filtered through a 70-μm cell strainer, and protease inhibitor and DNase were added before pelleting at 300g for 5 minutes. Cells were resuspended in the resuspension buffer consisted of Hibernate A (BrainBits, cat. no. HALF500), B27 (Thermo Fisher, cat. no. A3582801), and 10% FBS (Avantor, cat. no. 97068-085). Y-27632 ROCK inhibitor (Selleck Chemicals, cat. no. S1049) was added to all media used (dissection, digestion, resuspension) in 1:1000 ratio to reduce cell death during single-cell prep. Next, live RFP-expressing cells were sorted by FACS and processed using the 10x Genomics 3' v3.1 Dual Index workflow for scRNA-seq library generation by the UNC AA Core. gRNA reads were separately enriched from cDNA using custom-designed primers targeting the CROP-seq backbone sequence for assigning gRNAs to cells (GuideEnrich_1, GuideEnrich_2, GuideEnrich_3, and GuideEnrich_4; see Table S6 for primer sequences). The resulting libraries were sequenced with Illumina high-throughput se- quencers (NextSeq2000 P2 or P3 for gene expression library, NovaSeq6000 SP for gRNA expression library). We used cycles R1 x i1 x i2 x R2 = 28 x 10 x 10 x 90, with 1% PhiX for gene library, and cycles R1 x i1 x i2 x R2 = 28 x 10 x 10 x 42, with 20% PhiX for gRNA library.

QUANTIFICATION AND STATISTICAL ANALYSIS

MPRA data preprocessing

Each barcode DNA library was sequenced to an average read depth of 81M reads (51-109M). Each barcode RNA library was sequenced to an average read depth of 167M reads (155-191M). Since our MPRA library contained 35,891 oligo sequences (35,682 variants + 209 positive and negative controls), each oligo sequence was sequenced to an approximate read depth of ∼2,257 for DNA and ∼4,563 for RNA. Barcode-variant relationships were decoded from the barcode mapping step. From mapped barcodes, we generated barcode reference files and aligned our variant-barcode pairs using BWA.[68,52] From the alignment, we created RNA and DNA count matrices using featureCounts[69] (version 2.0.6) using the following parameters: –minOverlap = 20, –fracOverlap = 1, –fracOverlapFeature = 1. We excluded barcodes which were mapped to more than one variant. In total, we detected 397,152 barcodes that are uniquely mapped to each variant. RNA and DNA barcode counts were then aggregated for each variant. When applying quality control (QC) metrics, we first identified and filtered out outlier barcodes, defined as individual barcodes that had a $\log_2\left(\frac{RNA}{DNA}\right) > 2 + \text{median}(\log_2\left(\frac{RNA}{DNA}\right))$ for the given variant. This filtering step resulted in 397,055 barcodes, dropping out 97 barcode outliers. We also filtered out (1) variants with <5 barcodes per allele, and (2) variants which did not have measured risk

and protective allelic pairs. This resulted in 381,012 barcodes, with each variant mapped to an average of 11.86 barcodes (median value of 11). Finally, variants were removed if, for either allele, the proportion of total read counts (DNA + RNA) from individual barcodes that could be spuriously mapped by allowing mismatches was greater than 30%. See Figure S1A for the data analysis flow chart.

Identification of active elements

To identify active regulatory elements, we used the mpra (version 1.24.0) R package[58] with the following code:

```
mpra_set <- MPRASet(DNA = dna_matrix, RNA = rna_matrix, eid = element_ID, eseq = NULL, barcode = NULL)
design_matrix <- data.frame(intcpt = rep(1, replicate_number))
fit <- mpralm(object = mpra_set, design = design_matrix, aggregate = "sum", normalize = FALSE, model_type = "indep_groups", plot = TRUE)
tr <- treat(fit, lfc=log2(1.238016))
mpra_result <- topTreat(tr, coef = 1, number = Inf)
mpra_element = mpra_result %>%
    group_by(variant) %>%
    filter(P.Value == min(P.Value))
mpra_element$fdr = p.adjust(mpra_element$P.Value, "BH")
mpra_active = mpra_element[mpra_element$fdr<0.05 & mpra_element$logFC>0, ]
mpra_inactive = mpra_element[mpra_element$P.Value>0.1, ]
```

Briefly, an MPRAset object (mpra_set) was created using the RNA (rna_matrix) and DNA (dna_matrix) matrices (rows: barcodes assigned to each oligo sequence, columns: replicates) and the corresponding oligo sequence IDs (element_ID). The design_matrix was defined by including an intercept term for each replicate. We then performed MPRA analysis using the mpralm function, with the MPRASet object and the design matrix as inputs. The RNA/DNA ratio for each oligo sequence was compared to the average RNA/DNA ratio (1.238016) of the negative control with sample outliers removed (i.e., for each negative control sequence, RNA/DNA ratios were compared against 22 biological replicates, and replicates exceeding >2 standard deviations were subsequently removed), from which the statistical significance was calculated. The resulting MPRA summary statistics were obtained using the topTreat() function.

For each element, two oligo sequences (one with the A1 allele and the other with the A2 allele) were tested. Therefore, there is a possibility that (1) only one of the sequences was classified as active, and (2) the activity of two sequences is not entirely independent. To address this, we selected the minimal p value from each of the two sequences for calculating Benjamini-Hochberg adjusted FDR for each element. Elements were defined to be active (mpra_active) if they met the significance threshold of FDR<0.05 and $\log_2$ fold change (FC)>0. Inactive elements (mpra_inactive) were defined when both alleles had p>0.1, suggesting that both alleles do not show signs of regulatory activity.

Stratification of active elements based on transposable elements

We identified MPRA-active and MPRA-inactive elements that overlap with transposable elements using RepeatMasker[61] (version Open-3.0) downloaded from the UCSC genome build hg19. We surveyed three types of repetitive elements: ERV1 superfamily of long-terminal repeat (LTR) retrotransposons, non-LTR retrotransposons, and hAT superfamily of DNA transposons.[70] Non-LTR transposons were further categorized into Alu repeats (representing short interspersed elements, or SINE), L1 (representing long interspersed elements, or LINE), and SINE-VNTR-Alus (SVA). Elements that exhibit >50% sequence overlap with a specific transposable element were categorized as transposable elements. For example, active elements with >50% sequence overlap with Alu repeats were classified as Alu-active elements, while those without such overlap were classified as noAlu-active elements.

Given the enrichment of Alu repeats in active elements, we compared the overlap of active/inactive elements with Alu repeats. We ranked active and inactive elements based on the p values of their enhancer activity, grouping them into 10 bins (deciles). The top bins consist of elements with the most significant regulatory activity (lower MPRA p values and higher enhancer activity), while the bottom bins consist of elements with the least significant regulatory activity (higher MPRA p values and lower enhancer activity). Subsequently, the percentage of elements within each bin that overlap with Alu repeats (>50% sequence overlap) was calculated.

Epigenetic annotation of active elements

To characterize the epigenetic properties of active elements during neurodevelopment, we overlapped them with (1) scATAC-seq peaks from the developing telencephalon[11] and (2) EPRI anchors in HNPs.[14] The coordinates of inactive, active, active-Alu, and active-noAlu elements were converted from hg19 to hg38 using the liftOver() function in the rtracklayer Bioconductor package[62] (version 1.54.0). These coordinates were then overlaid against the ATAC-seq peaks in each cell type as well as EPRI anchors in HNPs using the findOverlaps() function of the GenomicRanges R package. We quantified the number of elements overlapping ATAC-seq peaks (or EPRI anchors) and utilized these counts to conduct enrichment analysis. To compare the overlap between active

and inactive elements, we conducted Fisher's exact test with the contingency table: (1) active (active/active-Alu/active-noAlu) elements that overlap with CREs (ATAC-seq peaks/EPRI anchors) in a given cell type, (2) active elements that do not overlap with CREs in a given cell type, (3) inactive elements that overlap with CREs in a given cell type, and (4) inactive elements that do not overlap with CREs in a given cell type.

We next quantified the read coverage of MPRA-active, active-Alu, active-noAlu elements using scATAC-seq bigwig files for each cell type. To account for variation in read depth, we normalized the resulting read coverage by the read depth of each cell type. We then calculated the average normalized read coverage for each type of active element (active, active-Alu, and active-noAlu) and cell type. Subsequently, we performed scale normalization across cell types to assess the cell-type-specific accessibility of each type of active element.

TF motif enrichment analysis

We used HOMER[56] (version 4.10.4) to investigate the TF motif contents of regulatory elements identified by MPRA (i.e., MPRA-inactive, MPRA-active, active-Alu, and active-noAlu). In particular, we used the findMotifsGenome.pl script within HOMER to identify enriched motifs in genomic regions of interest with default parameters using hg19 coordinates. This program returns TFBMs that are enriched in a given region of interest with their respective percentage of sequences harboring its respective TFBMs and statistics (e.g., p and q value). We filtered our result by q<0.05. To obtain enriched TFBMs, we searched ±250-bp around our regulatory elements.

Motif enrichment scores were calculated by using the following equation:

$$\textbf{\textit{Motif Enrichment Score}} = \textbf{\textit{log}}_2 \frac{\sum \Delta \% \textbf{ \textit{of sequence overlap}} * \textbf{\textit{number of motifs}}}{\sqrt{\textbf{\textit{number of elements}}}}$$

Δ % of sequence overlap is defined as the % of target sequence overlap (e.g., active, active-Alu, active-noAlu) minus the % of MPRA-inactive element overlap.

Identification of regulatory variants

To identify variants with regulatory activity, we also used the mpra (version 1.24.0) Bioconductor package[58] with the following code:

```
mpra_set <- MPRASet(DNA=dna_matrix, RNA=rna_matrix, eid=varid, eseq=NULL, barcode=NULL)
batch <- as.factor(c(rep(c(rep(1,10),rep(2,12)),2)))
alt <- c(rep(FALSE,repnum),rep(TRUE,repnum))
design_matrix <- model.matrix(~alt+batch)
samples = rep(1:repnum,2)
tr <- mpralm(object=mpra_set,
    design=design_matrix,
    plot=T,
    aggregate="none",
    normalize=T,
    block=samples,
    model_type="corr_groups")
mpra_result <- topTable(tr, coef = 2, number = Inf)
mpra_result$fdr = p.adjust(mpra_result$P.Value, "BH")
mpra_allelic = mpra_result[mpra_result$fdr<0.05, ]
mpra_nonallelic = mpra_result[mpra_result$P.Value>0.1, ]
```

Briefly, an MPRASet object (mpra_set) was created using the RNA (rna_matrix) and DNA (dna_matrix) matrices (rows: measured variants, columns: aggregated barcode counts for each A1 and A2 allele across samples). The design_matrix was defined by regressing with an indicator for one allele compared to the other ("A2" compared to "A1") and for batch. We then performed MPRA analysis using the mpralm function, with the MPRASet object and the design matrix as inputs. The resulting MPRA summary statistics (mpra_result) were obtained using the topTable function. P-values (P.Value) extracted from the mpra_result indicate statistical significance of allelic activity for each variant. These p values underwent multiple testing correction using the Benjamini-Hochberg adjusted FDR procedure. Variants were defined as having allelic regulatory activity (mpra_allelic) when they met the significance threshold of FDR<0.05. We then combined these variants with allelic regulatory activity with MPRA-active elements to define emVars. We also defined MPRA-nonallelic variants when their allelic regulatory activity significance was greater than p>0.1 (mpra_nonallelic) and did not overlap with MPRA-active elements. We defined emVars as variants that exhibit significant differential regulatory activity between risk and protective alleles (FDR<0.05) and are located in MPRA-active elements.

Stratification of variants based on mDisorder

The original cross-disorder GWAS reported mDisorder values that represent the number of disorders associated with each locus surpassing the GWS threshold.[5] Consequently, all emVars within a given locus were assigned to their corresponding mDisorder (the number of disorder) category. Variants were further stratified based on mDisorder, where mDisorder1 represents variants associated with only one disorder, and mDisorder3 represents variants associated with three or more disorders (up to eight disorders). Variants associated with three or more disorders (mDisorder3) are considered pleiotropic.

Epigenetic annotation of emVars

We characterized the epigenetic properties of emVars by overlapping them with scATAC-seq peaks from the developing brain.[11] Because scATAC-seq peaks were in hg38, we converted the coordinates of emVars from hg19 to hg38 using the liftOver() function in the rtracklayer Bioconductor package[62] (version 1.54.0). We quantified the read coverage of emVars using scATAC-seq bigwig files for each cell type. We extended the flanking region of the variants from 150-bp to 500-bp ($\pm$250-bp centered on the variants) to enhance the ATAC-seq signals for quantification. We normalized the resulting read coverage by the read depth of each cell type, calculated the average normalized read coverage for each cell type, and performed scale normalization across cell types to assess the cell-type-specific accessibility of emVars. We repeated the same analysis after stratifying emVars into mDisorder1 and mDisorder3 emVars. To assess the cell type specificity of (mDisorder1 and mDisorder3) emVars, we conducted a paired t test to compare scaled-normalized chromatin accessibility between excitatory neuronal lineage (early EN, dlEN, ulEN) vs. the remaining cell types (RG, IPC, IN:Prog, IN:MGE, IN:CGE, AstroOligo, Micro).

Defining genomic background SNPs

We defined the genomic background SNPs by obtaining variants with matched minor allele frequency (MAF) and LD with no nominal association (p>0.05). For each MPRA-tested variant, we randomly obtained 10 variants within the same chromosome with matched MAF ($\pm$10%) and the number of variants in LD (r^2>0.8). We obtained all variants with their respective MAF and LD if less than 10 variants were selected for a given MPRA-tested variant. MAF values and the number of LD SNPs were obtained using the garfield program[54] (version 2.0), as previously described.[17]

TFBM alteration enrichment analysis

We used the motifbreakR[57] Bioconductor package (version 2.10.2) to identify TFBMs altered by emVars. Following their vignette on the Bioconductor website, we subsetted the TF motif database by Hsapiens and excluded stamlabs. The resulting database included TF motifs from HOCOMOCOv10, HOCOMOCOv11, JASPAR_CORE, JASPAR_2014, jaspar2016, jaspar2018, and jaspar2022. Then we ran the following code to obtain TFs with binding motifs that were altered by emVars:

```
motifbreakR(..., filterp = TRUE, method = "ic", threshold = 1e-4, bkg = c(A = 0.25, C = 0.25, G = 0.25, T = 0.25))
```

The reference and alternative allele's *p* values were calculated by using:

```
calculatePvalue(..., granularity = 1e-5)
```

The output file was filtered by Refpvalue & Altpvalue < 0.05 and effect = "strong" to identify strong TF motif alterations.

To calculate the TFBM alteration enrichment for emVars, we also conducted motifbreakR analysis on genomic background SNPs and compared them to the number of TFBM alterations. Fisher's exact test was conducted to quantify enrichment using the contingency table: (1) emVars that alter TFBM1, (2) emVars that do not alter TFBM1, (3) genomic background variants that alter TFBM1, and (4) genomic background variants that do not alter TFBM1.

From this analysis, we identified TFs with nominal enrichment (p<0.05) relative to the genomic background. We repeated the same analysis for mDisorder1 and mDisorder3 emVars to identify transcriptional regulators for disorder-specific and pleiotropic variants, respectively. We identified PPIs among the TFs (i.e., emVar, mDisorder1, and mDisorder3 TFs) using the STRING database[65] (version 12.0) by inputting these TFs for "multiple proteins" analysis.

PPI network of TFBMs

mDisorder1 and mDisorder3 TFs were queried with the STRING database (version 12.0).[65] We used the full STRING network, visualizing medium-confidence (0.4 interaction score) edges with no other interactors besides themselves. To compute PPI connectivity, we downloaded the table listing reciprocal edges (i.e., gene A-gene B, gene B-gene A) from STRING. From this data we derived a dataframe listing the number of non-zero edges (PPI connectivity) each TF has and used the Anderson-Darling k-Sample test to test the difference between the two distributions (mDisorder1 TFs vs mDisorder3 TFs).

Statistical fine-mapping of emVars

To run fine-mapping on our loci we first used our 136 index SNPs and identified variants that overlap with each locus using the GenomicRanges R package. For all variants within a given GWS locus, we ran PLINK[60] (version 1.90b3.45) using 1000 Genomes Phase 3[71] chromosomal bam files, with the --r and --matrix parameters to produce the LD matrices. We also calculated z-scores for each variant as $log_2\left(\frac{Odds\ Ratio}{Standard\ Error}\right)$. We then used the resulting LD matrices and z-scores to run each fine-mapping algorithm with default parameters and a specification to detect one putative causal variant per locus.

For SuSiE (version 0.12.35), we ran the following code to obtain PIPs and credible sets:

```
susie_rss(z = zscore, R = LDmatrix, L = 1, n=727126)
```

For CAVIAR, we provided one file with the LD matrix, and one file with the rsids and their z-scores for each locus, and ran CAVIAR (version 2.2) with -c = 1.

For FINEMAP, we provided the LD matrix and file containing a datatable with the rsid, chromosome, position, allele1, allele2, MAF, beta value, and standard error for each variant in each locus. We then created a master file for each locus, which is an aggregate table of all the previously mentioned information within a locus, and is the final input for FINEMAP. We ran FINEMAP (version 1.4.2) using the parameters --sss and --n-causal-snps = 1.

To visualize the number of overlapping variants among emVars and fine-mapped variants from three statistical fine-mapping methods, we utilized the UpsetR[66] (version 1.4.0) R package. We also visualized their correlation by comparing the PIP values of all tested variants across three fine-mapping algorithms.

To calculate the enrichment of emVars compared to MPRA-nonallelic variants that pass PIP threshold (PIP_{thr}), we conducted two-sided Fisher's exact test with the contingency table for each credible set size: (1) emVars with $PIP>PIP_{thr}$, (2) MPRA-nonallelic variants with $PIP>PIP_{thr}$, (3) emVars with $PIP \leq PIP_{thr}$, and (4) MPRA-nonallelic variants with $PIP \leq PIP_{thr}$.

To compare the MPRA p values and abs(log_2FC) values between FM-emVars and nonFM-emVars, we applied progressive thresholding based on the credible set size and calculated the statistical significance by using the two-sided Wilcoxon rank sum test.

eQTL overlap

The bulk homogenate brain meta-eQTL dataset was obtained from de Klein et al.[27] We utilized their cortex meta-eQTL (n=6,523) filtered by FDR<0.05. Cell-type-specific eQTL datasets were obtained from Bryois et al.[28] (n=192) and the PsychENCODE Consortium[32] (layer 2-3 excitatory neurons [L2-3], n=346; layer 4 excitatory neurons [L4], n=319, layer 5 excitatory neurons [L5], n=333; layer 6 excitatory neurons [L6], n=328; LAMP5-expressing interneurons [LAMP5], n=223; SST-expressing interneurons [SST], n=278; VIP-expressing interneurons [VIP], n=294; Parvalbumin-expressing chandelier cells [ChC], n=313).[32] Since the dataset from Bryois et al. did not include adjusted p values, we calculated adjusted p values using the qvalue[72] (version 2.26.0) Bioconductor package. Both cell-type-specific eQTL datasets were filtered by q<0.05.

We overlapped our emVars with eQTL resources, aligning them by chromosome and position, or by rsid. We encountered some discrepancies in genome assembly between our MPRA (hg19) and eQTL datasets (either in hg19 and hg38). Consequently, we converted our MPRA into hg38 using the liftOver() function in the rtracklayer[62] (version 1.54.0) Bioconductor package. Variants that weren't converted into hg38 were further converted into hg38 by using the getBM() function in the biomaRt[51] (Version 2.58.0) Bioconductor package. We overlapped our MPRA data with eQTL datasets in hg38 based on chromosome and position. If chromosome and position weren't provided, we matched solely by rsid. Colocalization analysis was conducted between meta-eQTL and cross-disorder GWAS using the coloc[53] (version 5.1.0.1) R package with default parameters, as previously described.[73] The resulting loci with colocalization posterior probability greater than 0.6 (H4 PP>0.6) were retained.

For variant-level analysis, the proportion of overlap between MPRA and eQTLs was calculated by dividing the number of emVars that overlap with meta-eQTL by the total number of emVars. IDE overlap proportion was calculated by dividing the IDE variants (variants with MPRA $log_2FC>0$ & eQTL beta>0, and vice versa) by the number of eQTL-overlapping emVars. Coloc overlap proportion was calculated by dividing the number of colocalized loci by the number of loci with IDE variants.

Gene mapping for emVars

We used a three-tiered approach to assign emVars to their respective target genes. First, we overlapped emVars with meta-QTL to select IDE variants. We then queried eGenes of those IDE variants. Second, we employed the ABC model, as previously described.[74] We identified genes located within the 1Mb flanking regions of each cross-disorder locus. For each gene, we identified variants within the same 1Mb flanking regions that could potentially affect that gene. Subsequently, we calculated the ABC score (chromatin accessibility within a ±250-bp region around the variant x chromatin contact frequency between the variant and the gene) for each variant-gene pair. For a given gene, we gathered ABC scores of all variants located within the 1Mb flanking regions, and calculated Z-scores from the ABC scores. We then selected variant-gene pairs with Z>1.5. We also identified the gene promoter with the strongest chromatin contact frequency for emVars that were not mapped by this approach. Finally, we leveraged long-range chromatin loops[75] to link emVars to the distal target genes. Gene coordinates were based on Gencode v19 and promoters were defined as 2kb upstream to each transcription start site (TSS), as previously described.[76] Only protein-coding genes were used for gene mapping.

Stratification of genes based on mDisorder

After assigning emVars to their corresponding target genes, we translated variant-level mDisorder to gene-level mDisorder. In cases where the same gene was associated with emVars with different mDisorder values, the highest mDisorder value was assigned to the gene. Consequently, we categorized genes as mDisorder1 or mDisorder3 based on the mDisorder values of the variants associated with the respective variants. Therefore, mDisorder1 and mDisorder3 genes represent genes with disorder-specific and pleiotropic effects, respectively.

Definition of local background genes

We also compiled genes located near the cross-disorder loci as the local background null set. Because currently used approaches – the ABC model, long-range chromatin loops, and eQTLs – map variants to their target genes within the ±1Mb boundary, we identified genes located within ±1Mb of all GWAS variants tested in our study. We then removed genes mapped to MPRA-active elements and emVars from this set to define a null set of 1,708 local background genes. These genes showed comparable expression levels to emVar genes in HNPs.[77]

Biological pathway analysis

We identified biological pathways enriched for a given gene set using the R package gprofiler2 (version 0.2.3)[55] with the command line:

```
gost(geneset, organism="hsapiens", ordered_query=F, significant=T, user_threshold=0.1, correction_method="fdr", sources=c("GO","KEGG",
"REAC"), custom_bg=backgroundgenes)
```

Here, geneset includes (1) genes that were mapped to emVars, including those mapped to mDisorder1 and mDisorder3 emVars and (2) local background genes (genes located nearby cross-disorder loci).

Comparison between rare and common genes

We compared emVar genes (common genes, prioritized from common variation) with genes that harbor rare PTVs associated with AUT (185 AUT rare genes reported by Fu et al.[36]) and DD[37] (285 DD rare genes reported by Kaplanis et al.[37]). These rare genes were compared against emVar genes using all protein-coding genes as the total gene set for Fisher's exact test. Enrichment values and their statistical significance were determined using the contingency table: (1) emVar genes that are also AUT/DD rare genes, (2) emVar genes that are not AUT/DD rare genes, (3) non-emVar protein-coding genes that are also AUT/DD rare genes, and (4) non-emVar protein-coding genes that are not AUT/DD rare genes.

Chromatin accessibility and gene expression profiling across neurodevelopment

We first analyzed the chromatin accessibility profiles of mDisorder1 and mDisorder3 emVars in various cell types during neurodevelopment.[11] Our focus centered on cortical cells representing a developmental lineage of excitatory neurons, which include RG, IPC, earlyEN, dlEN, and ulEN. For each variant, we calculated the average chromatin accessibility within each cell type. The resulting normalized chromatin accessibility of mDisorder1 and mDisorder3 emVars were visualized using the pheatmap R package (version 1.0.12),[59] and variants were subsequently clustered based on their similarities in chromatin accessibility profiles, as shown in the dendrogram. Specifically, mDisorder1 emVars formed 9 clusters, while mDisorder3 emVars formed 11 clusters. Further, we calculated the average scaled chromatin accessibility of the variants within each cluster for individual cell types. Cell types with an average scaled accessibility greater than 0.5 were considered enriched for that particular cluster. When only one cell type displayed an average scaled accessibility above 0.5, it indicated that the variants within the cluster were specifically active in that particular

cell type. Conversely, if multiple cell types had an average scaled accessibility above 0.5, it indicated the variants within the cluster were active across multiple cell types. To quantify the findings, we counted the number of variants in clusters enriched in one cell type versus clusters enriched in more than one cell type for both mDisorder1 and mDisorder3 emVars. This quantification was followed by Fisher's exact test using the contingency table: (1) mDisorder3 emVars within clusters accessible in >1 cell types, (2) mDisorder3 emVars within clusters accessible in one cell type, (3) mDisorder1 emVars within clusters accessible in >1 cell types, and (4) mDisorder1 emVars within clusters accessible in one cell type.

We repeated the same analysis using the expression profiles of mDisorder1 and mDisorder3 genes in the same cell types during neurodevelopment.[11] Briefly, we calculated the average expression level within each cell type for each gene, which was visualized with the pheatmap R package (version 1.0.12).[59] Genes were further grouped into clusters based on their similarities in expression profiles within the dendrogram, in which mDisorder1 genes formed 11 clusters, while mDisorder3 genes formed 9 clusters. The average scaled expression values of the genes within each cluster were calculated for individual cell types, and cell types with an average scaled expression greater than 0.5 were considered to be enriched for that cluster. We quantified the number of genes in clusters enriched in one cell type versus clusters enriched in more than one cell type for both mDisorder1 and mDisorder3 genes. This quantification was followed by Fisher's exact test using the contingency table: (1) mDisorder3 genes within clusters expressed in >1 cell types, (2) mDisorder3 genes within clusters expressed in one cell type, (3) mDisorder1 genes within clusters expressed in >1 cell types, and (4) mDisorder1 genes within clusters expressed in one cell type.

Co-expression and PPI network analysis

We compared the "hubness" of mDisorder1 and mDisorder3 genes within two types of networks: co-expression networks and PPI networks. We first leveraged co-expression networks derived from transcriptomic profiles pertaining to cortical development.[33] For each gene, we retrieved kME values within the co-expression module to which the gene is assigned. These kME values indicate the centrality of a gene within its associated co-expression module. We compared the distribution of kME values between mDisorder1 and mDisorder3 genes.

Next, we delineated PPI networks among the mDisorder1 and mDisorder3 genes using the STRING database (version 12.0) via the "multiple proteins" analysis feature. Because *HIST* genes are clustered within the genome and form dense PPI networks, they were excluded from the PPI analysis to mitigate potential bias in PPI connectivity. We then compared the number of PPI interactions between mDisorder1 and mDisorder3 genes by Anderson-Darling k-Sample test.

Developmental expression trajectory

We surveyed developmental expression trajectories of mDisorder1 and mDisorder3 genes as previously reported.[8] Log_2-transformed expression values were centered to the mean expression level per sample. mDisorder1 and mDisorder3 genes were selected for each brain sample, and their average[39] centered expression values were calculated for each brain sample. LOESS smooth curves with 95% confidence bands were plotted for mDisorder1 and mDisorder3 genes across developmental stages.

CROP-seq data analysis in hiPSC-derived neurons

scRNA-seq and gRNA library sequencing files were processed using the cellranger[50] count (version 7.0.0) pipeline. The resulting gene x cell and gRNA x cell expression matrices were imported into a Seurat[63] (version 5.0.2) object, which was then utilized to analyze knockdown effects of emVar target sites for nearby genes. A total of 110,414 cells were obtained post-quality control from 5 experimental batches. For quality control, cells with >15% mitochondrial genes and cells with $\leq 1{,}250$ unique genes were filtered out. Association between perturbation (gRNA) and target genes were obtained by running negative binomial GLM analysis. First, to mitigate the potential confounding effects of gRNAs targeting other variants in the same locus, we defined control cells as those containing no gRNAs for that locus, and perturbed cells as those containing at least 3 read counts for any of the four gRNAs targeting the variant of interest within the locus. Second, target genes were defined as genes located within ±1.5Mb of the variant targeted by the gRNA. Third, we modeled the expression of target genes using the following formula, employing a variable perturbation identity to capture the gRNA effects and using other variables to adjust for the covariates' effects:

Gene expression ~ perturbation identity (control or perturbed) + number of gene reads + number of gRNA reads + batch + mitochondrial gene %

For each perturbation-gene pair, the GLM coefficient and p value were obtained. We then computed left-tailed (negative coefficient) p values, which were adjusted with the Benjamini-Hochberg FDR procedure. Genes with FDR< 0.05 were considered as hits.

Additionally, we employed a permutation analysis to ensure that the significant variant-gene pairs are not due to miscalibration or test statistic inflation. First, from all cells in the dataset, we randomly drew pseudo-control and pseudo-perturbed cells matching the

number of actual control and perturbed cells. With pseudo-control and pseudo-perturbed cells, we tested the association between each variant-gene pair using the same negative binomial GLM, computing the GLM coefficient for the perturbation. We repeated this process 2,000 times and derived 2,000 permuted coefficients. Finally, by comparing the actual coefficient (derived from real control and perturbed cells) to the distribution of the 2,000 permuted coefficients, we computed the one-sided permutation-adjusted p value. We confirmed that all of the significant emVar-gene pairs based on FDR had permutation-adjusted p value<0.05.

Finally, for the nearby gene down-regulation lollipop plots, we employed signed $-\log_{10} p$ values defined by: Signed $-\log_{10} p = -\log_{10} p_{\text{left-tailed}}$ x sign(GLM_{coef}), where $p_{\text{left-tailed}}$ and GLM_{coef} represent the left-tailed p value and GLM coefficient of a given variant-gene pair respectively. The results from this analysis, including each variant-gene pair's GLM coefficient, left-tailed p value, permuted p value, FDR, and signed $-\log_{10} p$ values are provided in Table S5.

Mouse CROP-seq data processing

Sequencing files were processed with the cellranger count pipeline (version 6.1.2) to output gene x cell and gRNA x cell matrices, which were imported into a Seurat (version 5.0.2) object for subsequent processing and analyses. Cells with ≥10% mitochondrial genes and cells with no gRNA reads detected were removed from the dataset. A total of 83,386 cells were obtained post-quality control from 10 experimental batches (15 mice). Unique molecular identifier (UMI) counts across different batches were log-normalized, then batches were integrated following Seurat standard procedures for UMAP visualization.

Differential gene expression analysis with mouse CROP-seq data

All 83,386 cells were used for differential gene expression analysis. Raw UMI counts were used for this analysis. For each CRISPR target gene, cells were classified as either control or perturbed based on gRNA expression. Control cells were defined as those with 0 gRNA reads for the target gene, while perturbed cells were those with at least 3 gRNA reads corresponding to either of the 2 gRNAs targeting the gene.

After grouping into control and perturbed cells, the expression of each gene in the transcriptome was modeled using a negative binomial GLM as described below:

Gene expression $\sim$ perturbation identity (control or perturbed) + number of gene reads + number of gRNA reads + batch + mitochondrial gene %

The resulting p values associated with the GLM coefficients for the perturbation identity (perturbed) were adjusted with the Benjamini-Hochberg procedure. Genes with FDR<0.05 were defined as DEGs. Next, we employed a permutation analysis to ensure that our called DEGs are not due to miscalibration or test statistic inflation. First, from all cells in the dataset, we randomly drew pseudo-control and pseudo-perturbed cells matching the number of actual control and perturbed cells. With pseudo-control and pseudo-perturbed cells, we tested the association between the perturbation and each DEG using the same negative binomial GLM, computing the GLM coefficient for the perturbation. We repeated this process 2,000 times and derived 2,000 coefficients. Finally, by comparing the actual coefficient (derived from real control and perturbed cells) to the distribution of the 2,000 permutation-derived coefficients, we computed the one-sided permutation-adjusted p value. A total of 46 *Anp32e* DEGs and 244 *Kmt5a* DEGs had a permutation-adjusted p value<0.05. We used these as final DEGs. Lists for the significant DEGs, all tested genes, their GLM coefficients and p values are provided in Table S5.

Cellular expression of mouse CROP-seq DEGs

For mouse CROP-seq data, cell types were annotated using the SingleR[64] (Version 2.2.0) package with the Allen Brain Mouse Whole Cortex & Hippocampus 10x dataset (2021)[78] as a reference. Cells were first labeled with subclasses from the Allen Brain metadata and then regrouped into six major classes: astrocyte, GABAergic neuron, glutamatergic neuron, microglia, oligodendrocyte, and vascular cells. Mean expression across the DEGs was calculated for each cell using log-normalized counts and visualized on UMAP which includes all 83,386 cells.

For fetal human brain data, we utilized single-cell gene expression data from Nowakowski et al.[48] DEGs were converted into human orthologs, then the mean expression values of DEGs were calculated for each cell type and visualized in a scaled heatmap.

Biological pathway analysis of mouse CROP-seq DEGs

We identified biological pathways enriched for CROP-seq DEGs using the R package gprofiler2 (Version 0.2.3) with the command line:

```
gost(degs, organism="mmusculus", ordered_query=F, significant=T, user_threshold=0.05, correction_method="fdr", sources=c("GO","KEGG", "REAC"), custom_bg=NULL)
```

We used gene ontology terms with 5<genes<800.

Expression of the human orthologs of mouse CROP-seq DEGs in postmortem brains of individuals with schizophrenia

Single-nucleus RNA sequencing data from the postmortem human prefrontal cortex, which includes control and schizophrenia subjects, was used.[49] Using this dataset, a 'DE score' was calculated for each gene in each cell type using the formula: DE score = $-\log_{10} p_{adj}$ x logFC, where p_{adj} and logFC represent the adjusted p value and log-fold change of a given gene in schizophrenia, respectively. Next, we employed a permutation approach to assess the extent to which CROP-seq DEGs are dysregulated in the human schizophrenic brain. For each cell type, we calculated the sum of DE scores for the CROP-seq DEGs (genes converted to human orthologs). We then randomly drew n genes (where n is the number of CROP-seq DEGs) from the intersection of genes expressed in both CROP-seq and human brain datasets and computed the sum of DE scores for those n genes. We repeated this process 10,000 times. By comparing the actual DEGs' sum of DE scores to the distribution of permuted scores, we computed a one-sided p value. If $p<0.05$, the CROP-seq DEGs are considered significantly increased or decreased in the brains of individuals with schizophrenia compared to random sets of genes. The signed $\log_{10}(p$ value) is visualized for each cell type to indicate whether the CROP-seq DEGs collectively exhibit increased or decreased expression.

PPI network of mouse CROP-seq DEGs

DEGs were queried with the STRING database (version 12.0).[65] We used the full STRING network, visualizing medium-confidence (0.4 interaction score) edges with a maximum of 10 interactors (1st shell). To compute PPI connectivity, we downloaded the table listing reciprocal edges (i.e., gene A-gene B, gene B-gene A) from STRING. From this data we derived a dataframe listing the number of non-zero edges (PPI connectivity) each DEG has, and visualized them in a density plot using non-zero edge nodes. Anderson-Darling k-Sample test was used to test the difference between the two distributions (*Anp32e* DEGs vs *Kmt5a* DEGs).
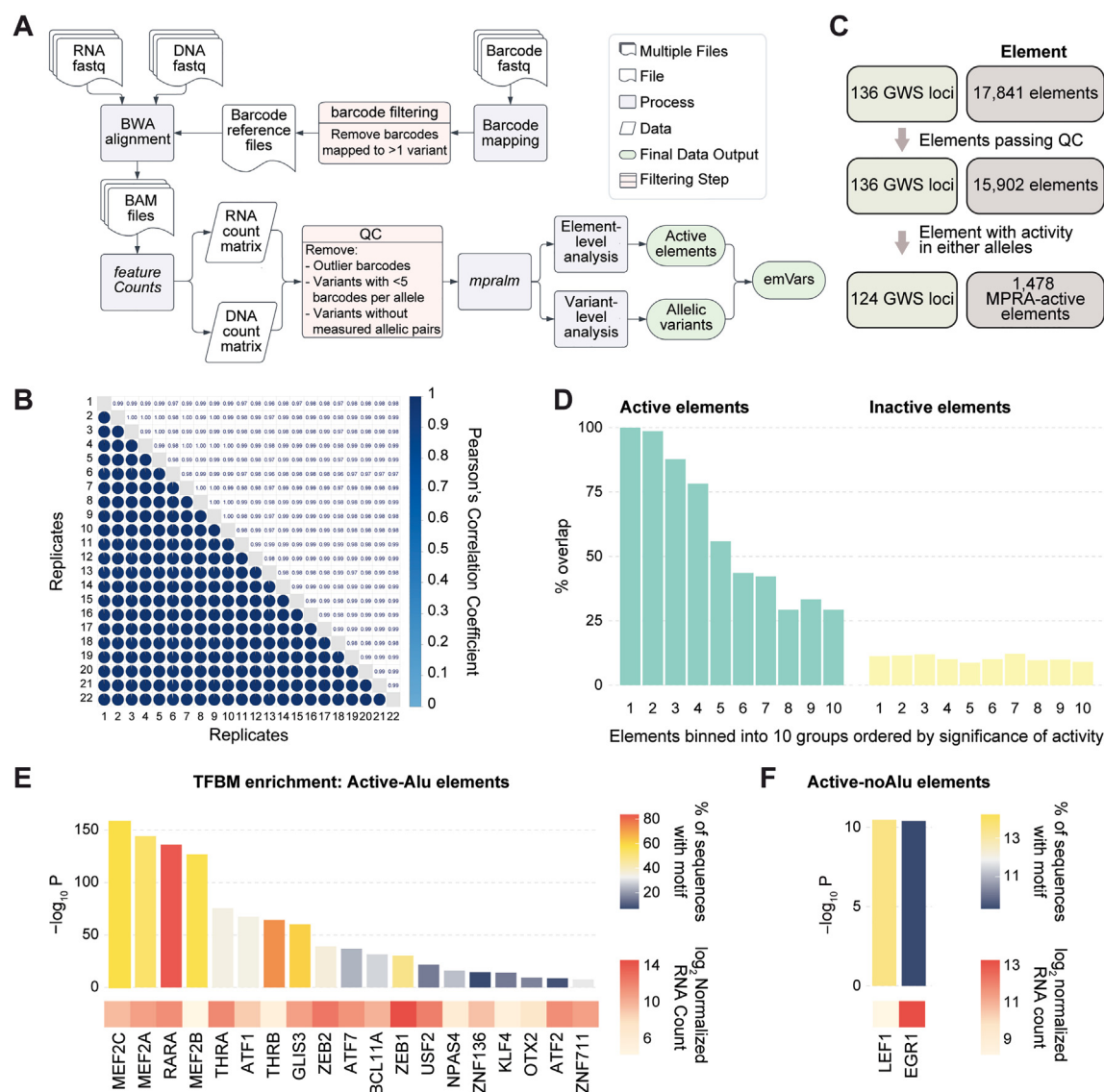
Supplemental figures

Figure S1. MPRA data processing and characterization of MPRA-active elements, related to Figure 1

(A) A flow chart schematic illustrating the steps utilized in the MPRA pre-processing pipeline.

(B) Reproducibility across 22 biological replicates was quantified using Pearson's correlation coefficients. Pearson's correlation coefficients for each pair of replicates are displayed as numerical values (top right) and as pie graphs (bottom left). The average r value (measured as RNA/DNA ratios between the biological replicates) was 0.985.

(C) A schematic of the number of MPRA-active elements in cross-disorder GWS loci. Out of 17,841 elements that harbor cross-disorder risk variants, 15,902 elements passed quality control (QC) and were thus tested via MPRA. We identified 1,478 elements across 124 GWS loci that show enhancer activity.

(D) Percent overlap of the MPRA-active and MPRA-inactive elements binned into 10 groups ordered by the significance of enhancer activity with Alu repeats.

(E and F) TFBM enrichment analysis and the TF expression level in HNPs for active-Alu (E) and active-noAlu (F) elements.
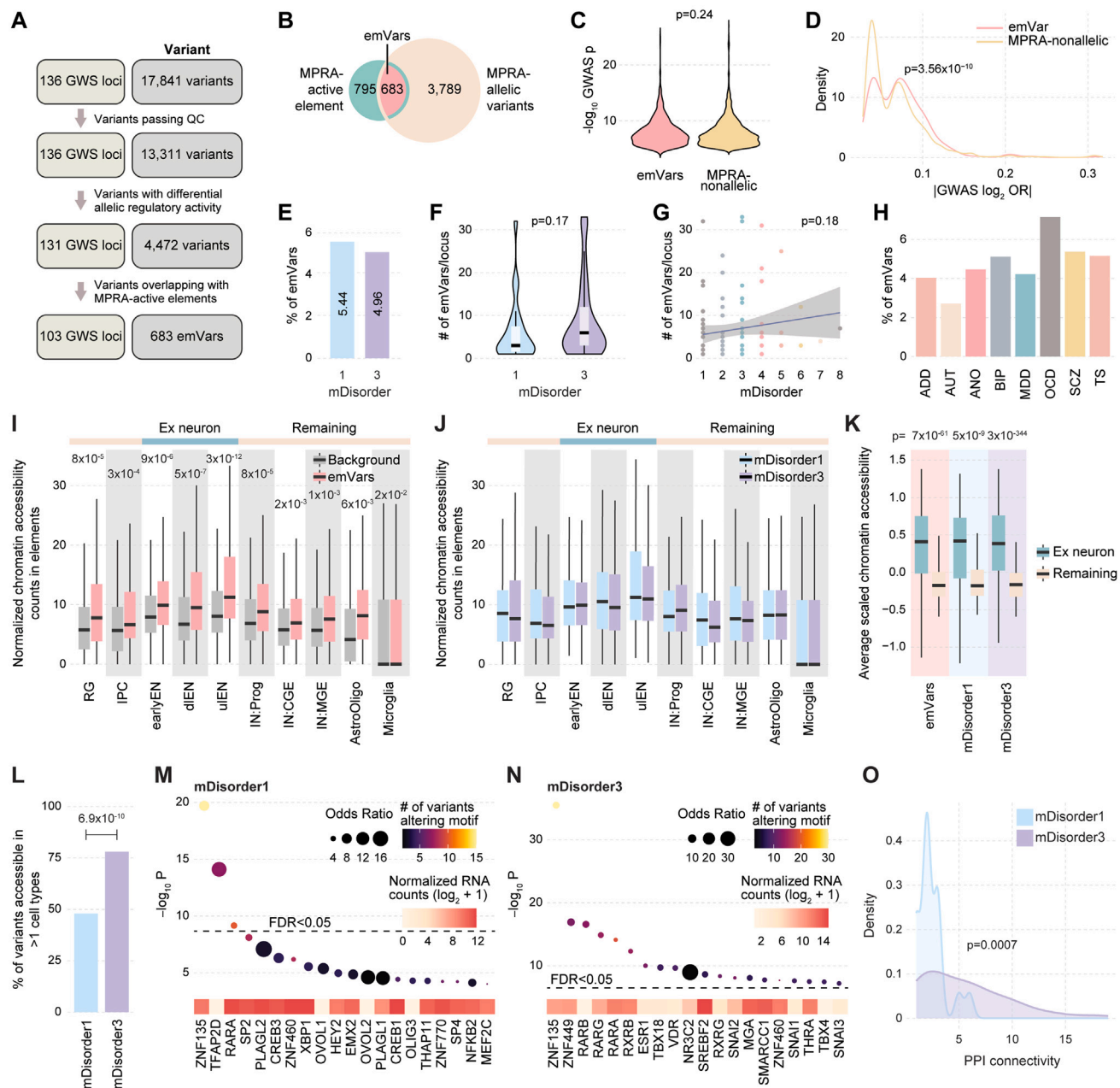
Figure S2. Identification and characterization of emVars, related to Figure 2

(A) A schematic of defining emVars. Out of 17,841 variants in 136 GWS loci, 13,311 variants passed QC, and their allelic regulatory activity was measured. We identified 4,472 variants with allelic activity across 131 GWS loci. We further stratified these allelic variants based on their overlap with MPRA-active elements, which resulted in 683 emVars across 103 GWS loci.

(B) emVars were defined as the intersection between MPRA-active elements and MPRA-allelic variants.

(C) Comparison of GWAS p values between emVars and MPRA-nonallelic variants. p value calculated by two-sided Wilcoxon rank sum test.

(D) Comparison of |GWAS log$_2$OR| between emVars and MPRA-nonallelic variants. p value calculated by two-sided Wilcoxon rank sum test.

(E) Percentage of emVars for mDisorder1 and mDisorder3 loci.

(F) The number of emVars per locus does not differ between mDisorder1 and mDisorder3 loci. p value calculated by two-sided Wilcoxon rank sum test.

(G) The number of emVars per locus stratified by mDisorder. p value calculated by linear regression.

(H) Percentage of emVars for each psychiatric disorder.

(I) Normalized chromatin accessibility counts of emVars and background variants in developing brain cell types. The cell types representing the excitatory neuronal lineage (Ex neurons) and the remaining cell types used for comparison in (K) are color-coded at the top. p values at the top of each boxplot indicate the significance of the difference in chromatin accessibility between emVars and background variants, calculated by the independent t test.

(J) Normalized chromatin accessibility counts of mDisorder1 and mDisorder3 emVars in developing brain cell types.
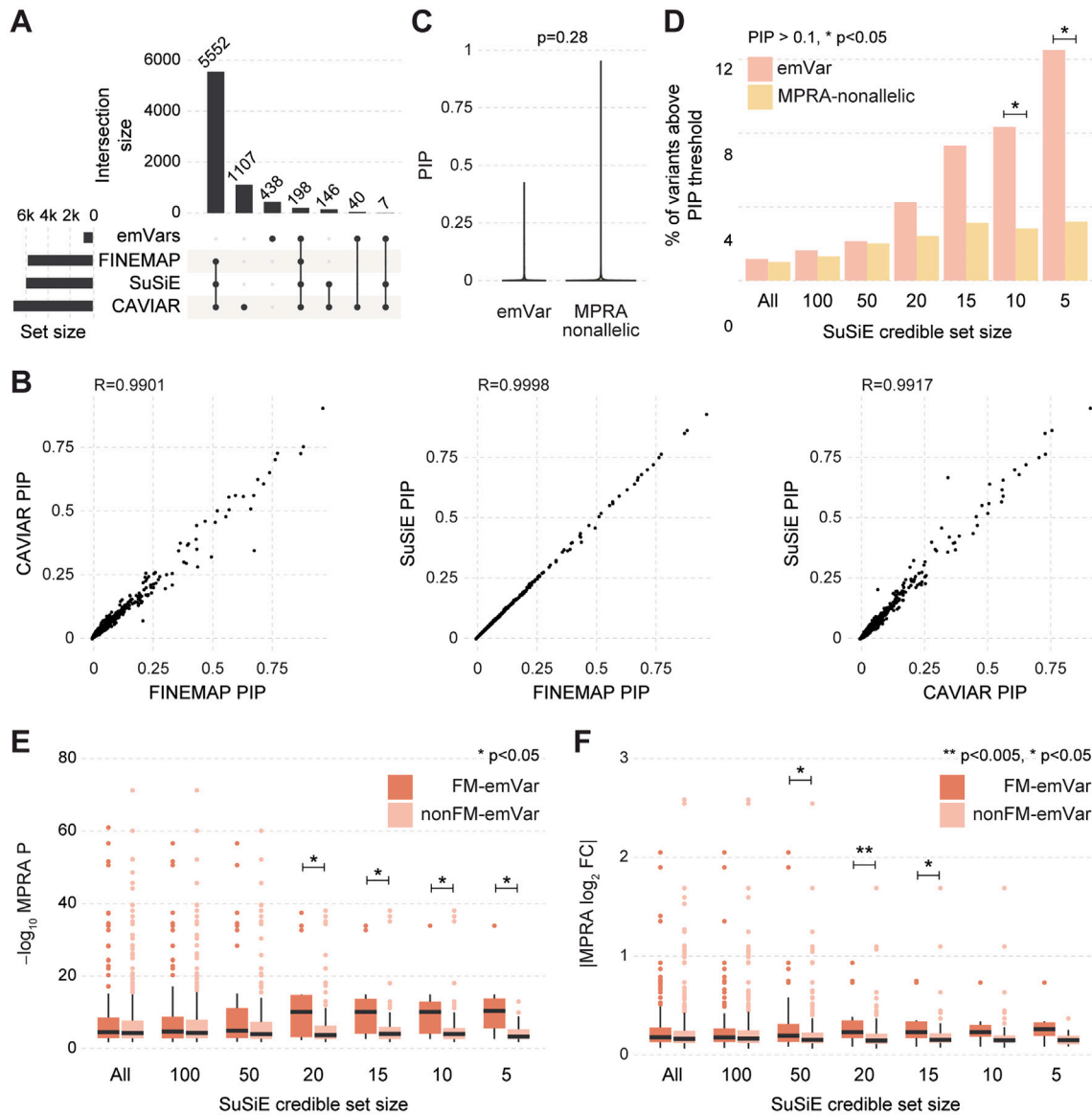
(legend continued on next page)

Figure S3. Statistical fine-mapping of emVars, related to Figure 2

(A) Overlap between emVars and credible SNPs identified by FINEMAP, SuSiE, and CAVIAR.

(B) Correlation of PIP values of all variants tested among three fine-mapping algorithms. R denotes the Pearson's correlation coefficient.

(C) PIP distribution of emVars and MPRA-nonallelic variants, $p = 0.28$. p value calculated by two-sided Wilcoxon rank sum test.

(D) Proportion of variants with PIP > 0.1 across different credible set sizes for SuSiE. p values calculated by two-sided Fisher's exact test.

(E and F) MPRA-allelic p values (E) and allelic effect sizes (F) between FM-emVars and nonFM-emVars across different credible set sizes. p values calculated by two-sided Wilcoxon rank sum test.
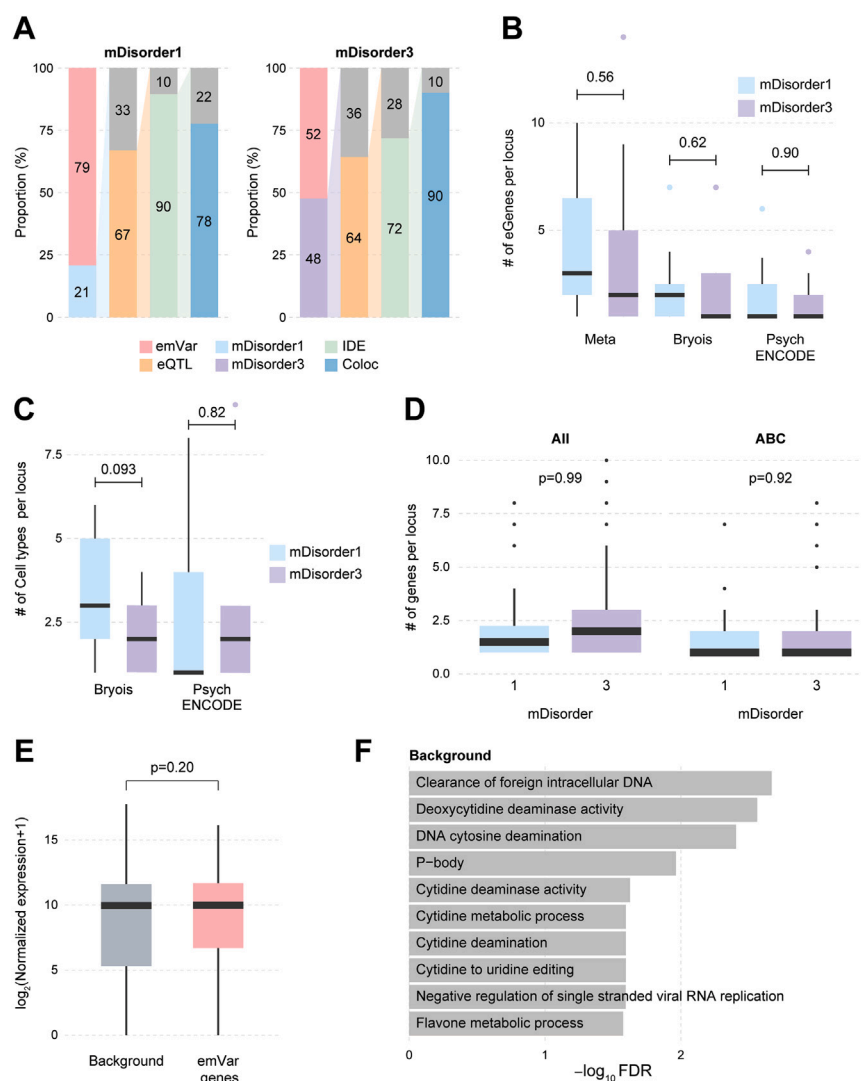
Figure S4. Characterization of genes mapped to emVars, related to Figure 3

(A) Overlap with meta-eQTL for mDisorder1 (left) and mDisorder3 (right) emVars. mDisorder1, percent of emVars located in mDisorder1 loci; mDisorder3, percent of emVars located in mDisorder3 loci; eQTL, percent of (mDisorder1/mDisorder3) emVars that overlap with eQTLs; IDE, percent of eQTL-overlapping emVars with IDE; Coloc, percent of loci where IDE emVars show colocalization with cross-disorder GWS loci.

(B) The number of eGenes for mDisorder1 and mDisorder3 loci based on meta-eQTL (Meta, left), Bryois et al. (middle), and PsychENCODE (right) datasets. *p* values calculated by two-sided Wilcoxon rank sum test.

(C) The number of cell types in which eGenes were detected for mDisorder1 and mDisorder3 loci using Bryois et al. and PsychENCODE eQTL resources. *p* values calculated by two-sided Wilcoxon rank sum test.

(D) The number of emVar genes for mDisorder1 and mDisorder3 loci when combining all three gene mapping approaches (All, left) and when using the ABC model (ABC, right). *p* values calculated by two-sided Wilcoxon rank sum test.

(E) Local background genes show similar expression levels to emVar genes in HNPs. *p* values calculated by two-sided Wilcoxon rank sum test.

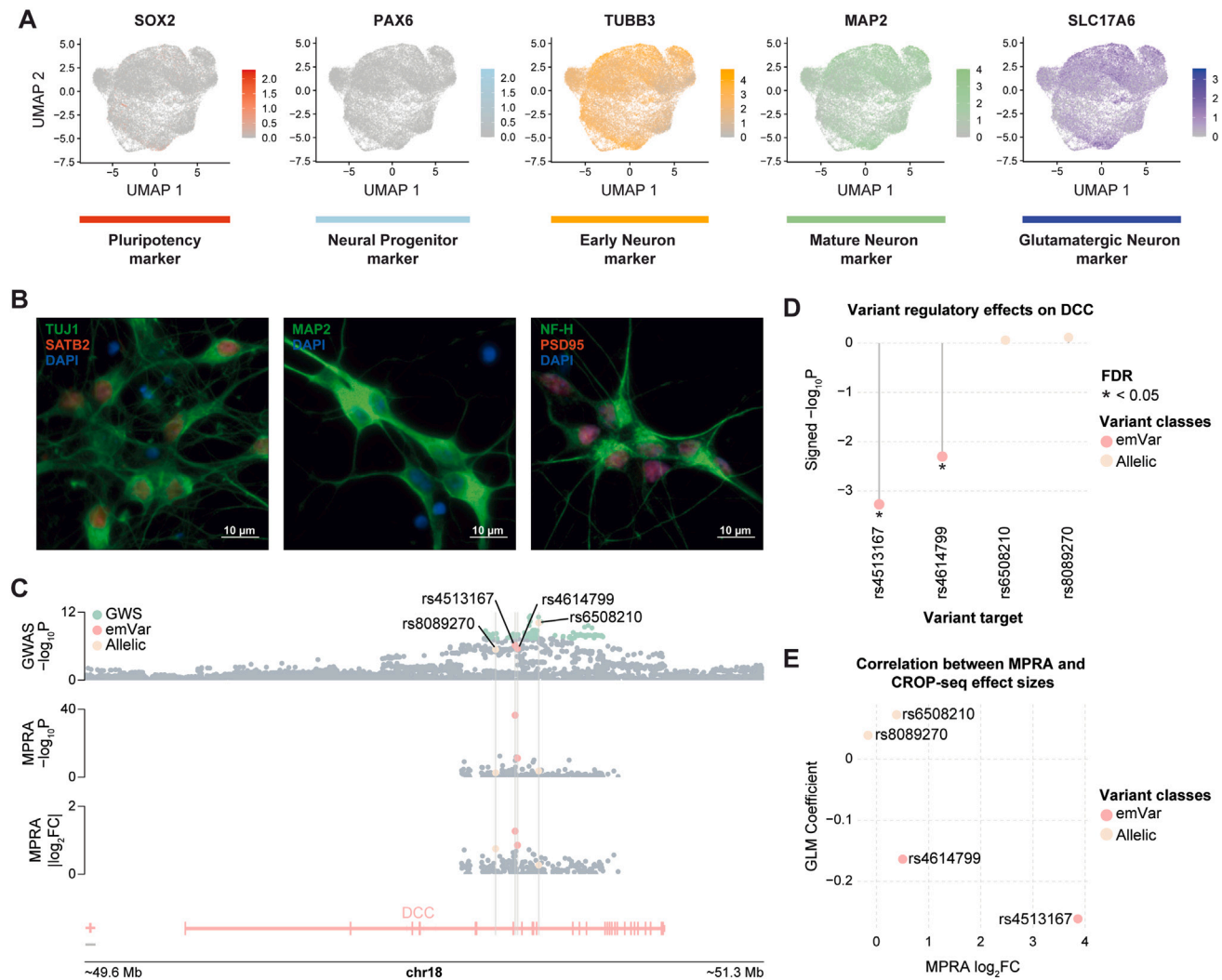(F) GO analysis on local background genes.

Figure S5. Transcriptomic impacts of emVars and MPRA-allelic/inactive variants in hiPSC-derived neurons, related to Figure 4

(A) Characterization of hiPSC-derived neurons derived from scRNA-seq data shows: no expression of pluripotency marker *SOX2* (red), no expression of neural progenitor marker *PAX6* (light blue), high expression of early neuronal marker *TUBB3* (yellow), high expression of mature neuronal marker *MAP2* (green), and high expression of glutamatergic neuronal marker *SLC17A6* (dark blue).

(B) Immunofluorescence staining of hiPSC-derived neurons on day 11 of differentiation reveals clear expression of the pan-neuronal markers TUJ1, MAP2, and NF-H (green). Additionally, the cortical neuronal marker SATB2 (red) and the excitatory neuronal marker PSD95 (red) are also prominently expressed.

(C) Locus plot of the *DCC* locus denoting locations and selection criteria for the two emVars (rs4513167 and rs4614799) and two MPRA-allelic/inactive variants (rs8089270 and rs6508210) labeled accordingly on the plot. Top panel shows GWAS association statistics of the emVars (pink), MPRA-allelic/inactive variants (pale yellow), and GWS variants (lime green). The second and third panels show significance and effect size of MPRA-allelic activity, respectively, for emVars (pink) and MPRA-allelic/inactive variants (pale yellow). Bottom panel shows genes within a 1.7Mb window around the target variants. Gray bars highlight variant locations along the gene track.

(D) Lollipop plot shows perturbation at the rs4513167 and rs4614799 emVar target sites led to significant downregulation of DCC in comparison to rs8089270 and rs6508210 MPRA-allelic/inactive target sites. Signed $-\log_{10}p$ was plotted for each variant in the DCC locus to represent both the direction of the effect as well as significance, with FDR-adjusted p value significance being denoted by an asterisk.

(E) Correlation between element-level MPRA $\log_2$FC and CROP-seq effect size (GLM coefficient) for CROP-seq tested emVars and MPRA-allelic variants.
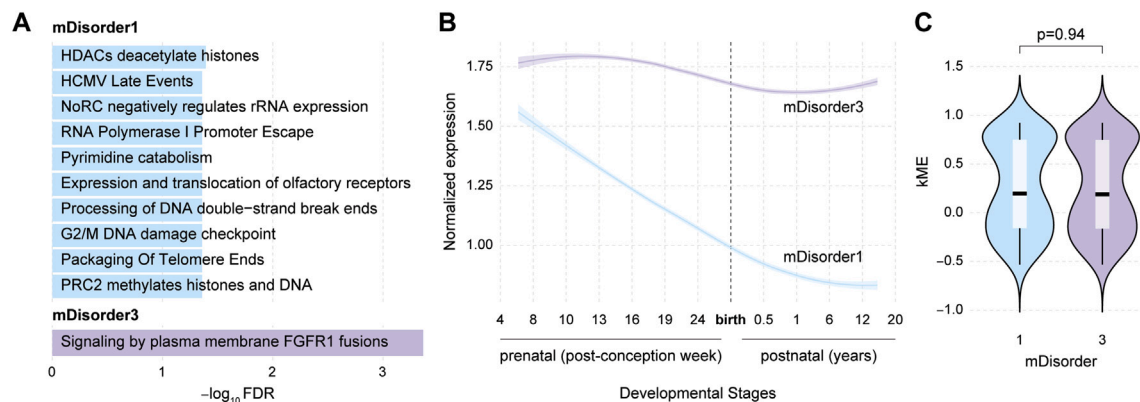
Figure S6. Gene Ontology, developmental trajectory, and co-expression network of genes mapped to mDisorder1 and mDisorder3 emVars, related to Figure 5

(A) GO analysis on mDisorder1 (top) and mDisorder3 (bottom) genes.

(B) Developmental expression trajectories (LOESS smooth curves) of mDisorder1 and mDisorder3 genes.

(C) Distribution of kME values for mDisorder1 and mDisorder3 genes. *p* value calculated by two-sided Wilcoxon rank sum test.