nature genetics

Article

https://doi.org/10.1038/s41588-023-01533-5

Integrative analyses highlight functional regulatory variants associated with neuropsychiatric diseases

Received: 24 June 2021

Accepted: 15 September 2023

Published online: 19 October 2023

Check for updates

Margaret G. Guo[1,2], David L. Reynolds[2], Cheen E. Ang[3,4,5], Yingfei Liu[5,6], Yang Zhao[2], Laura K. H. Donohue[2,7], Zurab Siprashvili[2], Xue Yang[2,8], Yongjin Yoo[5], Smarajit Mondal[2], Audrey Hong[2], Jessica Kain[7], Lindsey Meservey[9], Tania Fabo[2,7], Ibtihal Elfaki[2,7], Laura N. Kellman[2,8], Nathan S. Abell[7], Yash Pershad[4], Vafa Bayat[10], Payam Etminani[10], Mark Holodniy[11,12], Daniel H. Geschwind[13], Stephen B. Montgomery[3,7], Laramie E. Duncan[14], Alexander E. Urban[7,14], Russ B. Altman[1,4,7], Marius Wernig[3,5] & Paul A. Khavari[2,8,15] ✉

Noncoding variants of presumed regulatory function contribute to the heritability of neuropsychiatric disease. A total of 2,221 noncoding variants connected to risk for ten neuropsychiatric disorders, including autism spectrum disorder, attention deficit hyperactivity disorder, bipolar disorder, borderline personality disorder, major depression, generalized anxiety disorder, panic disorder, post-traumatic stress disorder, obsessive-compulsive disorder and schizophrenia, were studied in developing human neural cells. Integrating epigenomic and transcriptomic data with massively parallel reporter assays identified differentially-active single-nucleotide variants (daSNVs) in specific neural cell types. Expression-gene mapping, network analyses and chromatin looping nominated candidate disease-relevant target genes modulated by these daSNVs. Follow-up integration of daSNV gene editing with clinical cohort analyses suggested that magnesium transport dysfunction may increase neuropsychiatric disease risk and indicated that common genetic pathomechanisms may mediate specific symptoms that are shared across multiple neuropsychiatric diseases.

Genome-wide association studies (GWAS) of neuropsychiatric disorders have identified thousands of risk loci, most of which reside in noncoding DNA of possible regulatory function in neural cell types[1–5]. Decoding mechanisms whereby such variants mediate disease risk have proved difficult. First, most GWAS-identified single-nucleotide polymorphisms, or variants (SNVs), are in linkage disequilibrium (LD) with other adjacent SNVs[6], underscoring the need for single-nucleotide resolution studies of allele-specific activity. Second, most GWAS variants reside in noncoding DNA regions distal to brain-expressed coding genes[7], making variant–gene linkage challenging[8]. Finally, many neuropsychiatric disorders may arise, in part, from dysfunctions during brain development[9], emphasizing the need to assess variants in models of human neural differentiation. A single-nucleotide resolution resource of risk variants with altered transcription-directing activity in human neural cell types may nominate dysregulated target genes to features of neuropsychiatric disease[10].

A full list of affiliations appears at the end of the paper. ✉e-mail: khavari@stanford.edu

Here we present a single-nucleotide resolution compendium of functional, differentially-active SNVs (daSNVs) in regulatory DNA associated with ten human neuropsychiatric diseases along with an analytical framework to assess how these variants dysregulate pathways implicated in disease. SNVs were identified from GWAS of autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), bipolar disorder (BPD), borderline personality disorder (BLPD), major depressive disorder (MDD), generalized anxiety disorder (GAD), panic disorder (PD), post-traumatic stress disorder (PTSD), obsessive-compulsive disorder (OCD) and schizophrenia (SCZ). Massively parallel reported assays (MPRA) compared activity in matched risk versus reference allele pairs in embryonic stem (ES) cells, ES cell-derived human neural progenitor cells (NPC), differentiating neurons, and glial cells. MPRA identified 892 daSNVs with differing transcription-driving functions from nonrisk counterpart alleles, many with neural cell-state specific differential activity. Transcriptomic, chromatin accessibility and looping data from human neural cell types were generated and integrated with expression quantitative trait loci (eQTL) data to link daSNVs to dysregulated target genes (eGenes). The resulting 641 putative eGenes contain multiple known and new drug targets within protein interaction networks implicated in neurologic function. Integration of daSNV–eGene-pathway connectivity with population-scale genotyping-clinical data nominated pathomechanisms for specific neuropsychiatric symptoms. These data provide a single-nucleotide resolution resource of functional variants linked to common human neuropsychiatric disorders and provide a framework to connect these variants to shared pathomechanisms, therapeutics and clinical features.

Results

Active regulatory variants linked to disease

To identify neuropsychiatric daSNVs that alter transcription-directing activity, MPRA[11,12] was used. A total of 2,221 variant pairs were prioritized by first curating neuropsychiatric GWAS-indexed SNVs and clinically annotated disease-relevant SNVs (n = 15,904 initial SNVs), expanding them to all SNVs in LD ($r^2 > 0.8$) and then filtering through DNase I hypersensitive data from ENCODE in 40 different neural and immune cell types (Fig. 1, Fig. 2a and Supplementary Table 1). Genotype-tissue expression (GTEx) eQTL datasets[13] were used to annotate potential target genes for these 2,221 variants. On average, five SNVs were tested per locus. SNVs tested reflected the genomic associations currently available. For instance, neuropsychiatric GWAS datasets are most extensive for SCZ, and thus roughly half of the tested variants were associated with this disorder. Additionally, roughly 12% of the SNVs tested were associated with more than one disease, indicating possible shared genetic signals across multiple neuropsychiatric disorders.

SNV function was assayed via MPRA in human H9 ES cells and their differentiated anterior and posterior neural stem cell progeny along with neurogenin-2-induced neuronal cells undergoing a previously defined maturation time course[14] to mimic features of neural development along with human astrocytes (AST), and multiple cell lines, including the HEK293T epithelial kidney line used widely for MPRA, medulloblastoma cells lines D283 and D341 and neuroblastoma cell lines IMR-32 and SH-SY5Y. Per-replicate concordance in cell-extracted barcode codes from the resulting 44 independent MPRAs indicated robust data quality (Extended Data Fig. 1). A total of 892 daSNVs (false discovery rate (FDR) < 0.05) with significant allelic activity differences in any of the cell types tested were found using a generalized linear model[15]. The validity of the model was confirmed with QQ plotting, which confirmed RNA counts followed a negative binomial distribution, whereas plasmid counts approximate a Poisson distribution (Extended Data Fig. 1e). Approximately 30% of variants were significant per disease; distribution of cell type significance remained similar across diseases, with ~14.5% SNVs significant within each cell type and condition. HEK293Ts displayed a different pattern of allele specificity from other neural cell lines and the ES-derived neural cells (Fig. 2a–e). A likelihood ratio F-test model determined that 326 of 892 daSNVs had significant interaction terms between cell type and allelic activity. Of the 187 daSNVs with significant cell-specific activity in human neural cells, only 72 overlap with those in HEK293T cells (Fig. 2f). GO term analysis of the nearest genes showed that ES-derived neural cell types and neural cell lines capture more neural-specific processes, such as 'postsynapse organization' or 'neural projection development', compared to HEK293Ts. Additionally, neural cell types displayed strong cell-type-dependent allele-specific activity in immunomodulation. A cell-type-specific positive correlation was noted between the neuropsychiatric MPRA dataset and noncoding sequence-based variant predictive model, DeepSea[16] (Fig. 2g and Supplementary Table 2). Neuropsychiatric genetic variants have cell-type-specific activity at a DNA base pair scale not uncovered using conventional HEK293T cells.

RNA and chromatin dynamics in neural differentiation

To place daSNVs in genomic context, matched RNA-seq, chromatin accessibility profiling via ATAC–seq[17] and enhancer–promoter looping via H3K27ac HiChIP[18] were performed in ES cells at days 2 (N-D2), 10 (N-D10) and 28 (N-D28) of neuronal differentiation, as well as anterior (A-NPC) and posterior (P-NPC) neuronal progenitor cells along with primary human AST from adult brains (Figs. 1b and 3 and Extended Data Fig. 2). Neurodevelopmental markers were observed in a cell-type-specific fashion (Fig. 3a). AST markers (*GFAP, PDGFRA* and *S100B*) were preferentially expressed in primary AST as was a glutamatergic specific marker (*SLC1A1*) in induced neurons. Anterior marker *FOXG1* and posterior marker *HOXA2* were seen in A-NPC cells and P-NPC cells, respectively. GABAergic markers demonstrated lower expression in induced neurons, consistent with a primarily excitatory phenotype. Motif scanning of differential ATAC-accessibility profiles found transcription factor (TF) motifs clustered in the following two groups: early neuronal differentiation motifs (*FOXB1, ZIC1* and *NEUROD1*) and later neuronal differentiation motifs (*CUX1* and *ONECUT1*; Fig. 3b,c). ATAC footprinting captured TF temporal dynamics (Extended Data Fig. 2). Processed H3K27ac HiChIP data extracted unique and shared regulatory DNA loops (Extended Data Fig. 2d,e and see Supplementary Methods for more details). The majority of putative enhancers were linked to promoters distal to the nearest gene (Extended Data Fig. 2c), as previously found[19]. Comparing these epigenomic datasets to published studies[19–21] found strong overlap in accessibility profiles but less comparability for looping data due to differences in experimental methods (Supplementary Tables 3 and 4). Additionally, a comparative analysis between our daSNVs and allele-specific open chromatin hits in ref. 20 found overlapping significant variants (n = 206) show a moderate degree of positive correlation in log-fold changes (r^2 = 0.48, $P = 1.7 \times 10^{-13}$; Extended Data Fig. 2g). Transcriptomic and epigenomic data provided an integrative view of *cis*-transcription regulatory modules and verified the phenotypes of neural cell populations studied.

The regulatory effects of neuropsychiatric disease variants were next explored. First, enrichment of the 892 MPRA-identified daSNVs within differential H3K27ac loops and accessible regions was quantified for each cell type (Fig. 3d and Supplementary Table 5). daSNVs for ADHD, BPD, PD, GAD and SCZ were enriched in ES cells as well as in A-NPCs and P-NPCs, with no enrichment seen for control variants associated with type 2 diabetes risk. daSNVs were also not enriched in AST differential loops and accessible regions, suggesting that neuropsychiatric disease pathogenesis may unfold primarily in neuronal cells. Second, allele-specific ATAC and H3K27ac HiChIP analysis found that daSNVs were infrequently associated with changes in DNA accessibility (2% of daSNVs) or looping (1% of daSNVs), suggesting that tested daSNVs do not mediate large changes in chromatin accessibility and structure. A total of 268 daSNVs (30.0%) directly altered at least one known TF motif while 239 (26.8%) of daSNVs created at least one new TF

https://doi.org/10.1038/s41588-023-01533-5

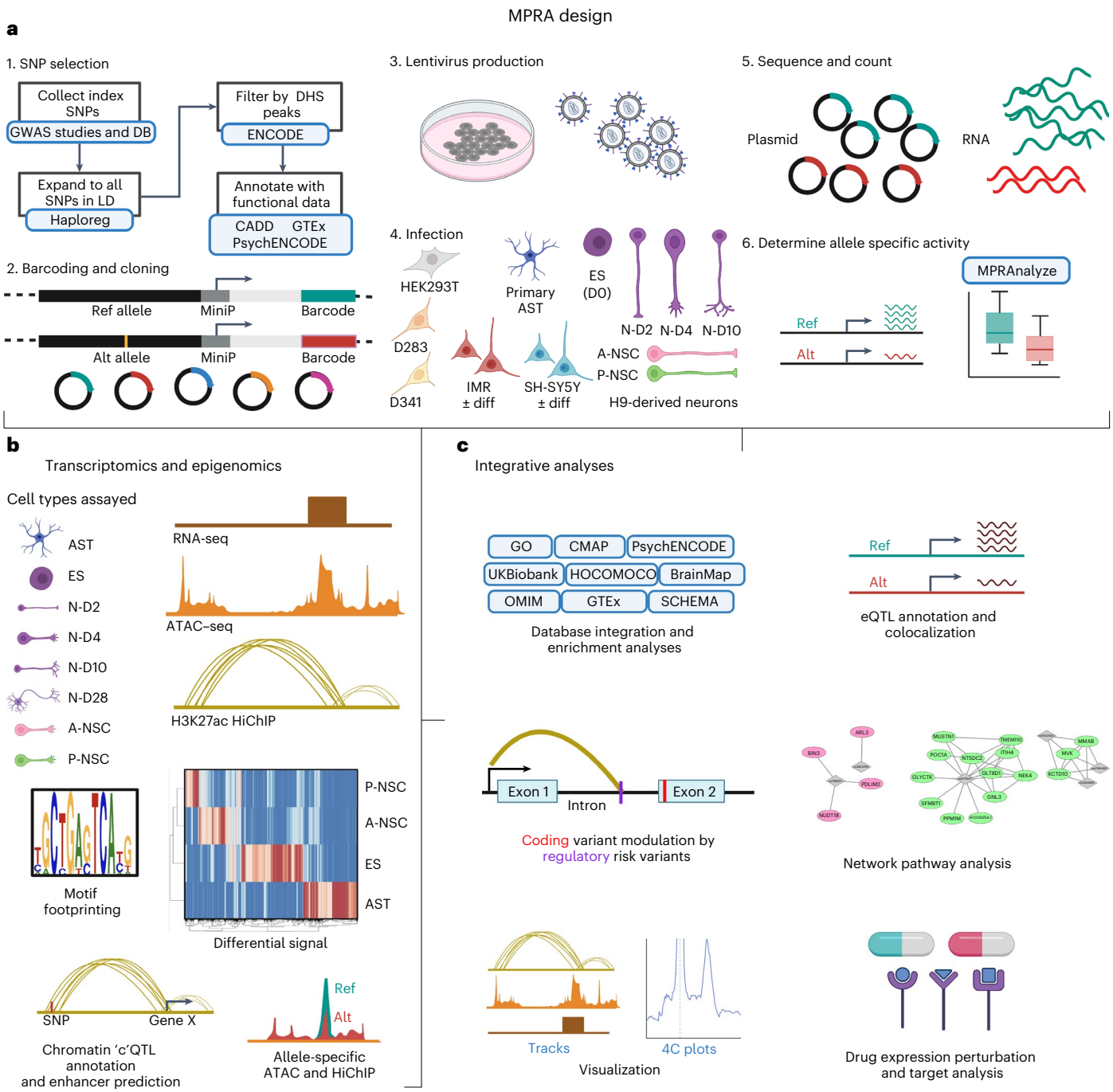


Fig. 1 | MPRA, transcriptomic and epigenomic integrative analyses for neuropsychiatric disorders. a, Schematic depicting lentiviral MPRA design for uncovering allelic-specific activity. A total of 250 bp oligos were designed to assay 2,221 neuropsychiatric disease GWAS loci. SNVs were selected from GWAS studies and then filtered through publicly available epigenomics and eQTL datasets. H9 human ES cells, H9-derived neurons on days 2, 4 and 10 of differentiation (N-D2, N-D4 and N-D10), A-NSC and P-NSC, respectively, AST and cell lines (HEK293Ts, D283, D341, IMR-32 cells (±differentiation), and SH-SY5Y (±differentiation) were infected with the lentiviral MPRA library. **b**, Schematic indicating the cell types in which RNA-seq, ATAC–seq and HiChIP were performed as well as subsequent assay-specific analyses. Heatmap shows differential RNA-seq expression. **c**, Analyses that integrate MPRA, transcriptomic and epigenomic data to explore transcription regulatory pathomechanisms. A-NSC, anterior neural stem cells; AST, astrocytes; P-NSC, posterior neural stem cells.

motif (Supplementary Table 6). Interestingly, motif sequences for TFs that influence neuronal or glial differentiation, including AP-1 motifs[22], SOX17 (ref. 23) and GLI1 (ref. 24), were found substantially broken or gained at daSNVs in a disease-specific manner (Extended Data Fig. 2f and Supplementary Table 6b), suggesting that daSNVs act through altering local neural TF action rather than via larger scale impacts on chromatin architecture.

To ensure that the epigenomic and transcriptomic profiles used in these analyses are substantially enriched for regions capturing the genetic heritability[25–27] of neuropsychiatric disorders, linkage disequilibrium score regression (LDSC)[28] was used to generate cell-specific heritability estimates using 25 neuropsychiatric GWAS. ATAC–seq-based cell-specific annotations were derived using open chromatin profiles for ten neural cell types. Cell-specific heritability estimates were generated

https://doi.org/10.1038/s41588-023-01533-5

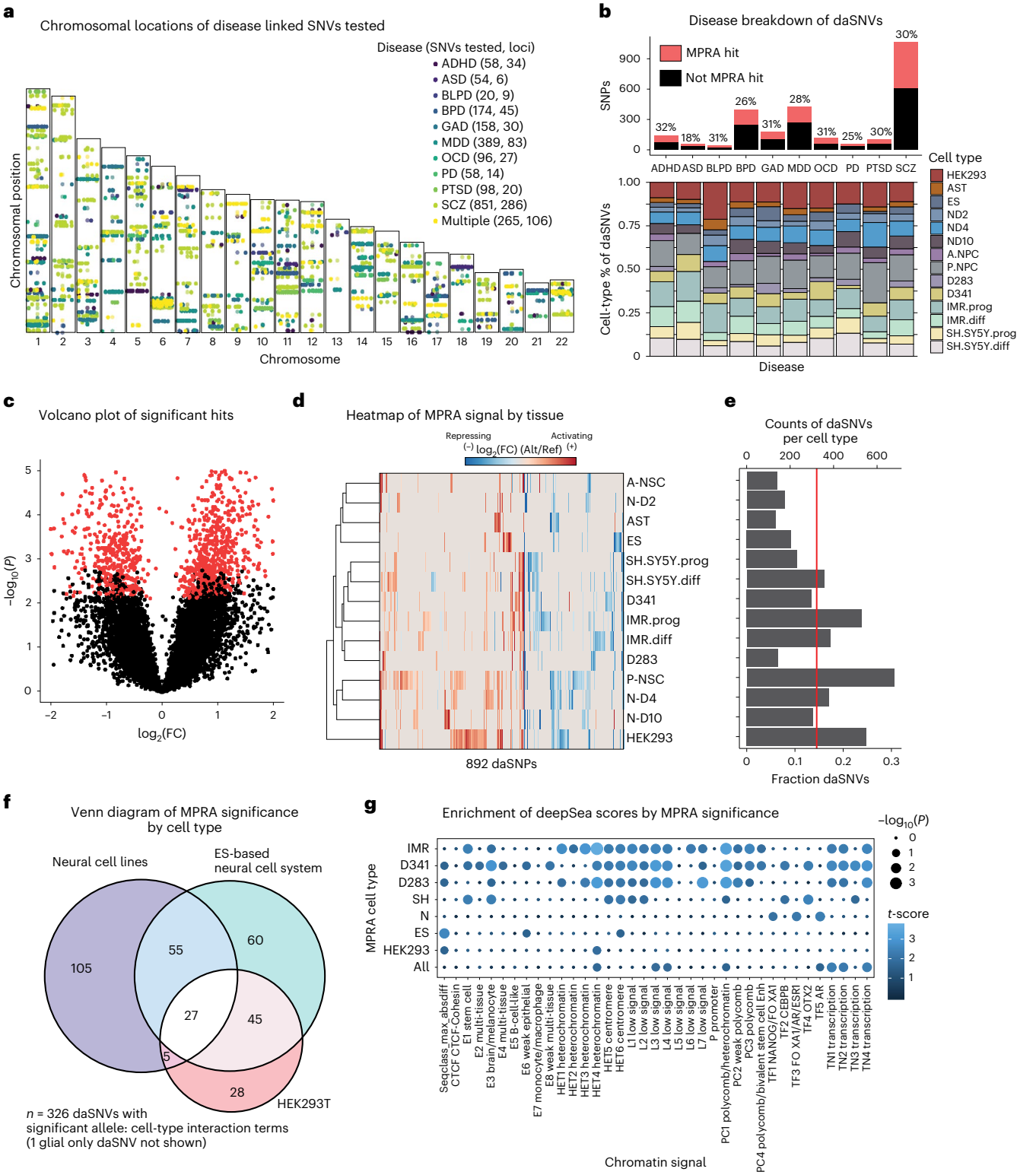
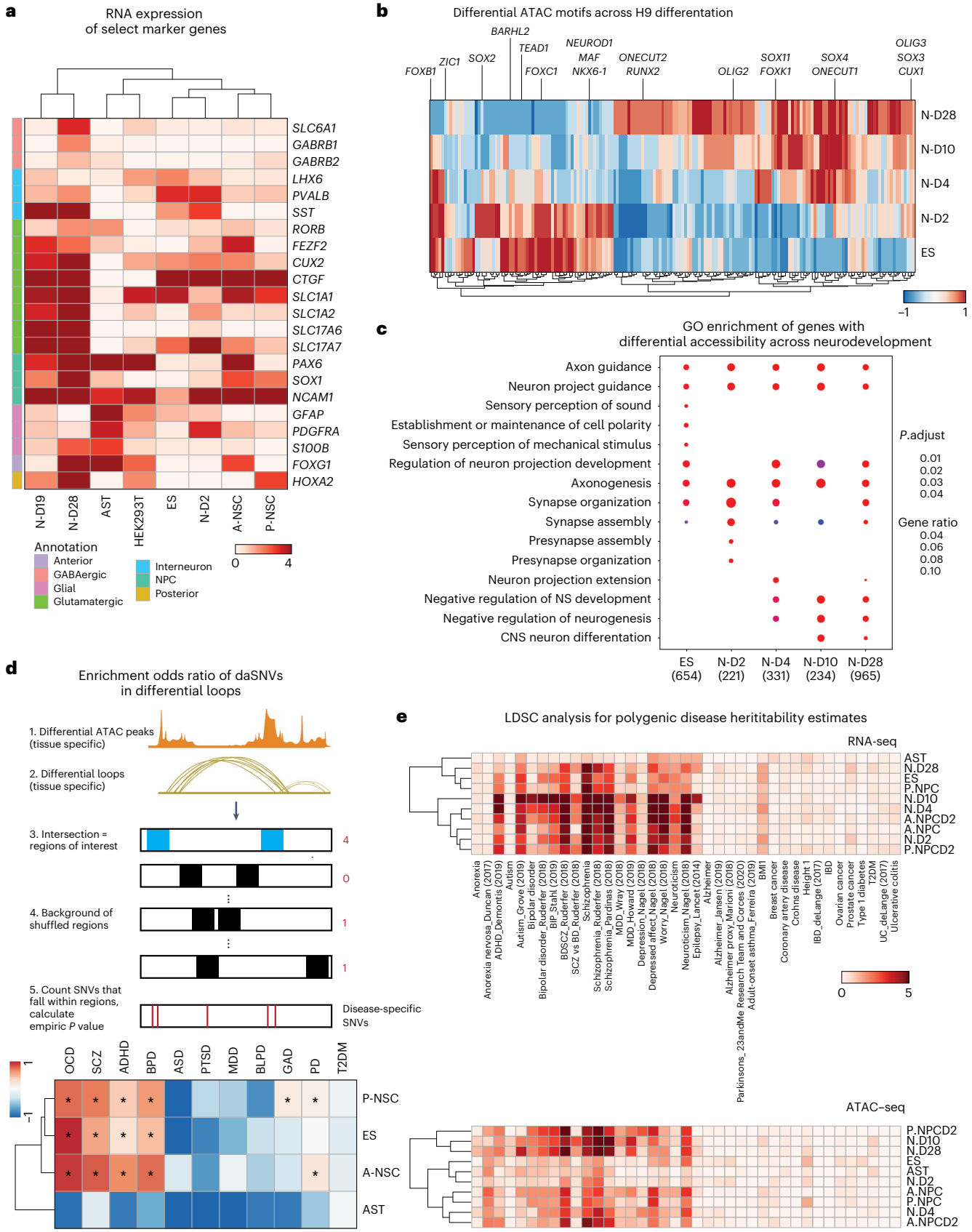


Fig. 2 | MPRA identifies 892 functional daSNVs across ten different neuropsychiatric diseases. a, Chromosomal map of locations of 2,221 SNPs tested and their disease annotations. **b**, Barplot (top) indicating the fraction of assayed SNPs that were significant, separated by disease; ~30% of SNVs tested were deemed significant, with the exception of ASD, barplot (bottom) shows further distribution of daSNVs across cell types and conditions tested. **c**, Volcano plot of $-\log_{10}(P)$ versus $\log_2$(FC) indicating significant hits (red dots). **d**, Heatmap of $\log_2$(FC) of alternative over reference allele activity captured by MPRA for the 892 significant hits. Left, a row-based dendrogram of the heatmap shows relatedness of cell types and conditions by daSNV profile. **e**, Right, barplot showing counts and fractions of daSNVs by cell type, the red line shows the average fractions of daSNVs significant by cell type = 0.145. **f**, Venn diagram of the daSNVs showing significant cell-type-dependent allelic activity (FDR < 0.05) within neural cell lines, the ES-based neural cell system and HEK293T (325 of the 326 significant variants are shown, as glial cells are not shown). **g**, Dotplot showing enrichment of DeepSea Scores based on MPRA significance, where color is the $-\log_{10}(P)$ and the size is the *t*-statistic for a two-sided Student *t* test between the distribution of the allelic differential for the sequences scores classes for daSNVs versus non-daSNVs. FC, fold change.

https://doi.org/10.1038/s41588-023-01533-5



for neuropsychiatric diseases with ATAC–seq-derived cell-specific annotations for 34 non-neural epithelial cells and cancer cell lines, as negative controls (Fig. 3e and Supplementary Data 1). At Bonferroni-corrected $P < 0.01$ with respect to multiple hypothesis testing for each of the 44 cell types, the epigenomic profiles of NPCs and later-stage neurons were enriched for regions capturing the heritability of numerous neuropsychiatric GWAS [1,25,88–100], particularly for BPD, SCZ, MDD and neuroticism, unlike embryonic cells, AST, epithelial and cancer cell lines. RNA-seq-based

Fig. 3 | Epigenetic profiling of neural cell system shows neuropsychiatric disease relevance. a, RNA-seq expression heatmap of key marker genes for each cell type. *FOXG1* and *HOXA2* are expressed in our anterior and posterior neural stem cells, respectively. *PDGFRA* and *GFAP* are expressed in AST. SLC glutamatergic gene markers are expressed at later stages of neuronal differentiation. **b**, Heatmap of normalized motif occurrences in differential ATAC peaks over the time course of H9-derived neuronal differentiation. Motifs known to be more prominent in early neuronal differentiation (*NEUROD1* and *ZIC1*) are highlighted versus later neuronal differentiation (*SOX11* and *CUX1*). **c**, GO biological process enrichment dotplot showing Benjamin–Hochberg-corrected *P* values from two-sided hypergeometric tests for enrichment of genes found nearest to ATAC peaks within H9-derived cells across the temporal neuronal differentiation axis. **d**, Top, a schematic of the GWAS tissue and disease-specific enrichment approach used (see Methods for more details) to derive a heatmap (bottom) of enrichment odds ratio of the daSNVs by disease over differential loop regions that were filtered by ATAC peaks. T2DM was used as a control and indicated no enrichment. Enrichment was concentrated to the neuronal stem cells and the ES cell neuronal lineages (Supplementary Data 1). **e**, Cell-type-specific LDSC hereditability estimate negative $\log_{10}(P)$ heatmaps for RNA-seq (top) and ATAC–seq (bottom) in the ES-derived neural cell system and relevant neural cell types with GWAS extracted from ref. 88–102, and if not specified from ref. 25. NS, not significant; T2DM, type 2 diabetes mellitus.

cell-specific annotations were derived using gene expression profiles for ten neural cell types and cell-specific heritability estimates of neuropsychiatric diseases derived with RNA-seq-derived cell-specific annotations for 16 non-neural epithelial cells and cancer cell lines, as negative controls. At Bonferroni-corrected $P < 0.01$ with respect to multiple hypothesis testing for each of the 26 cell types, the transcriptomic profiles of embryonic cells, NPCs and neurons were enriched for regions capturing the heritability of numerous neuropsychiatric GWAS, particularly for ADHD, ASD, BPD, SCZ, MDD, worry and neuroticism, unlike the AST, epithelial and cancer cell lines. For additional negative controls, LDSC generated epigenomic and transcriptomic-based cell-specific heritability estimates using 19 GWAS for non-neuropsychiatric, complex diseases, none of which were significant. These analyses suggest the neural cells and corresponding open chromatin epigenomic and transcriptomic data used here are relevant cell states to neuropsychiatric disease heritability.

Linking regulatory variants to genes and cell types

Brain-specific GTEx[13] and PsychENCODE[2,29] eQTL data were next used to link the 892 identified daSNVs to genes whose RNA expression varies with the variant's presence (termed eGenes). 480 of the 892 daSNVs (approximately 67% of both the tested and significant SNVs) had eQTL–gene linkages where the gene was expressed in the neural cell model (transcripts per million (TPM) > 1), with 641 putative eGenes identified. Overlaying these 641 genes onto StringDB-nominated protein interaction networks for each disease (Supplementary Data 2). For SCZ, 447 genes across all cellular conditions were linked to eQTLs, annotated by protein function and disease-relevant automated literature search. A substantial portion of SCZ-linked candidate eGenes was associated with the following five main biological processes: ion channel activity, general neural processes, immune regulation, cell cycle regulation and transcription regulation (Fig. 4a), providing a resource for hypothesis generation, components of which also touch on major proposed SCZ pathomechanisms[1,30,31]. For example, the immune hypothesis of SCZ[30] was captured through genes connected to *C4A* and *HLA-C* loci. eGenes involved in protocadherin function, Notch signaling[32] and synaptic action via *SV2A*[33], were also observed. Network analyses also identified potential pathomechanisms. For instance, *SUFU* maintains neuronal identity during corticogenesis[34]; however, its putative role in SCZ has not been defined. Of interest, there was no significant linkage between target gene biological processes and cell-type-specific daSNV networks. BPD was linked to 176 putative eGenes, MDD to 128, ADHD to 67, GAD to 30, OCD to 27, BLPD to 24 and PTSD to 17 (Extended Data Fig. 3). MPRA-allele-specific activity was confirmed via both episomal and lentiviral luciferase assays for selected daSNVs linked to target genes of interest (*RERE*, *C4A*/*HLA-C*, *GIT1* and *PNMT*; Supplementary Table 7). Less than ten eGenes were nominated for ASD and PD; small numbers of daSNVs and putative eGenes correlated with fewer GWAS studies for these disorders. Putative eGene networks in MDD, GAD, OCD and ADHD, however, consistently captured features of the same five biologic processes seen in SCZ, suggesting dysregulation of common shared biologic processes underpins multiple neuropsychiatric disorders.

Transcriptional dysregulation, which can exert major impacts on neurodevelopment and homeostasis[35], is one such process notably enriched in MPRA-significant networks versus networks extracted from nonsignificant MPRA hits. Furthermore, *cis*-target genes of daSNV-linked TFs may serve as potentiators for further downstream pathogenic effects on myelination, brain development and membrane depolarization, among others (Fig. 4b). For example, four SCZ/BPD-linked daSNVs looped to the promoter for the *POU5F1*/*OCT4* TF, which establishes a pluripotency-neural differentiation axis with *SOX2* and *PAX6* (ref. 36); each of these daSNVs is a brain-specific eQTL. These daSNVs displayed different activities across neuronal differentiation. rs2442722 and rs35735140 showed allele-specific activity early, in contrast to later differences seen with rs28428768 and rs3134944 (Extended Data Fig. 4), suggesting *POU5F1*/*OCT4* dysregulation at different points in neuronal differentiation may confer differing levels of pathogenic risk. Studying daSNV activity dynamics in neuronal differentiation may shed light on stage-specific pathogenic mechanisms of transcriptional dysregulation.

To help place the 641 putative eGene targets of neuropsychiatric daSNVs in the human brain context, integrative analyses were performed with GTEx and single-cell RNA-seq data from the Allen Brain Atlas[37]. When daSNVs' colocalization signals were stratified by GTEx tissue type, brain tissue types had substantially higher signal than non-brain tissue types, indicating daSNVs preferentially modulate transcription within the brain (Supplementary Table 8). Next, single-cell RNA-seq of cortical neurons was used to match putative eGenes to cell subtypes. PCA analysis across the 127 neuron expression profiles placed genes on a spectrum with the following two main considerations: number of cell types where a given gene was expressed and whether the gene was primarily found in GABAergic or glutaminergic cells. Putative daSNV eGenes expressed in more subtypes (that is *CNNM2*, *ANK3* and *RTN1*) may be involved in more global neuronal processes, while certain

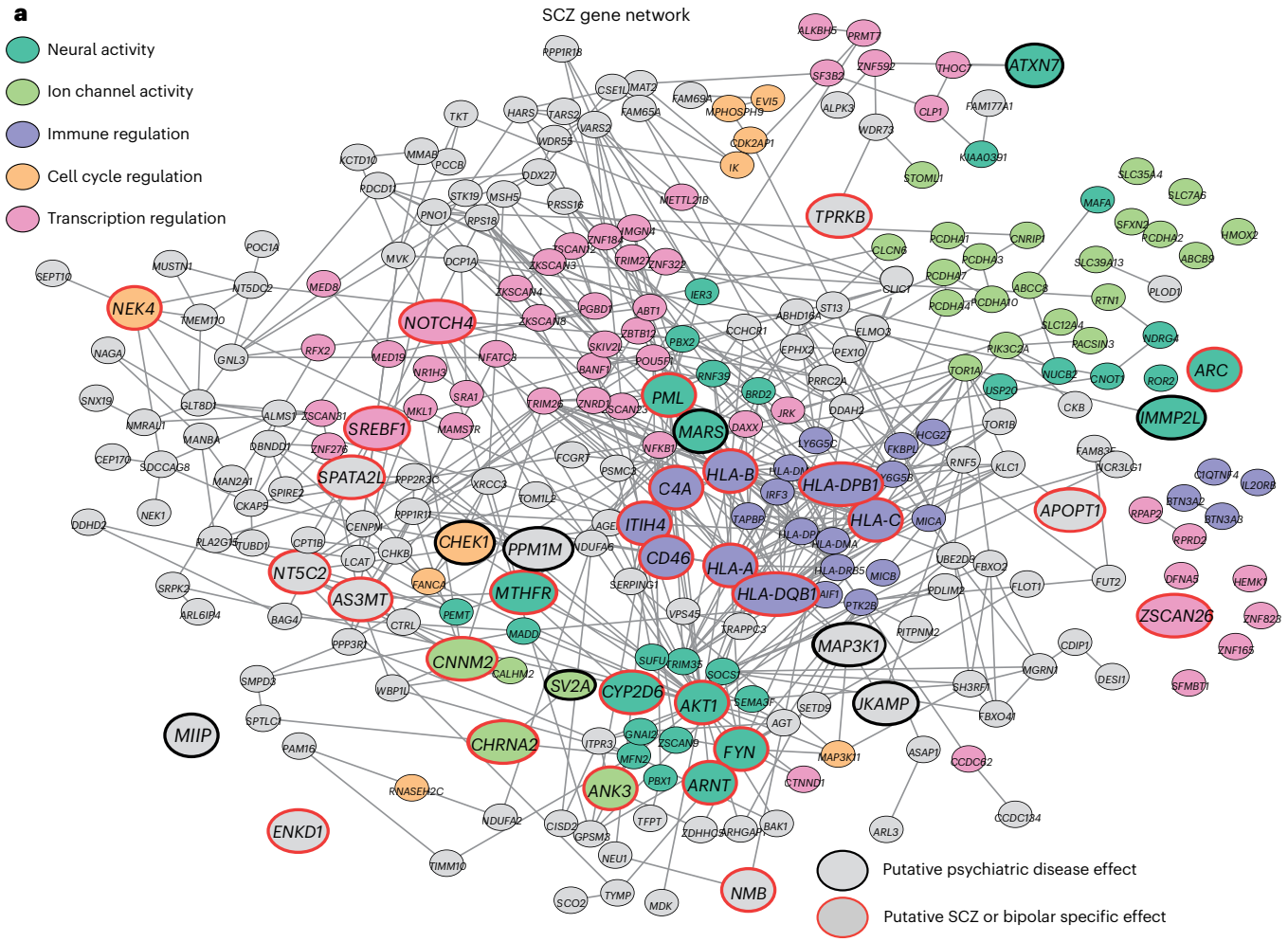
Fig. 4 | daSNV–eGene networks and their transcription regulatory effects. a, Protein–protein interaction network of eGenes assigned to SCZ-associated daSNVs by GTEx and PsychENCODE. Black outlined nodes were genes implicated in neuropsychiatric disorder disease risk based on an automated PubMed annotation pipeline; red outlined nodes were genes implicated specifically in SCZ. Nodes are color-coded by neuropsychiatric-relevant functional process. **b**, A dotplot depicting GO biological processes for target genes of TF daSNVs associated with neuropsychiatric diseases, where the color indicates FDR-adjusted two-sided *P* value from hypergeometric test for enrichment, the size of the dot indicated gene set size and is accompanied by an expression heatmap (left) showing $\log_{10}(\text{TPM} + 1)$ expression values for the TFs in neural relevant tissues assayed. **c**, Scatter plot of PC loadings for PC2 versus PC1, where loadings represent expression profiles from 127 cortical subtypes derived from Allen Brain Atlas scRNA-seq data, each point is an eGene. PC1 loadings correlate to expression of the gene in an increasing number of scRNA cortical subtypes. PC2 denotes the GABAergic versus glutaminergic cell type axis with *CHRNA2* having a mostly GABAergic signature, while *PTK2B* has mostly glutaminergic signature. PC, principal component.

https://doi.org/10.1038/s41588-023-01533-5

genes were primarily expressed in GABAergic cells (*CHRNA2* (ref. 38)) or in glutaminergic cells (*PTK2B*[39]; Fig. 4c and Supplementary Table 9). scRNA and GTEx-based analyses map daSNV candidate eGenes to cell types within human brain.

Refining target genes and the CNNM2 magnesium transporter

daSNVs' impact on local gene regulation was next studied by integrating eQTL analyses with chromatin accessibility and looping data. Focused analyses were necessary because an average of five daSNVs linked to



https://doi.org/10.1038/s41588-023-01533-5

Table 1 | daSNVs in neuropsychiatric diseases

daSNP	Lead SNP	OR/β	Diseases	MPRA cell type	daSNP avg *P* value	daSNP avg log_2(FC)	HiChIP genes	eGenes
rs3814546	rs1182	1.06	ADHD|BPD|MDD|SCZ	N	1.4×10^{-3}	0.57	–	*C9orf78|TOR1A|TOR1B|USP20*
rs2276834	rs7618915, rs11717383, rs353547	1.06	BPD|SCZ	HEK293T	4.6×10^{-5}	1.10	–	*DNAH1|GLYCTK|GNL3|ITIH4|MUSTN1|NEK4|NT5DC2|PPM1M|SFMBT1|TMEM110*
rs5996094	rs1023500	1.08	SCZ	N	3.6×10^{-4}	−0.77	–	*CCDC134|CYP2D6|DESI1|NAGA|TNFRSF13C*
rs139784051	rs8044995	1.08	SCZ	N	5.7×10^{-3}	0.53	–	*CTRL|ELMO3|ENKD1|GFOD2|LCAT|NFATC3|PLA2G15|PRMT7|SLC12A4|TSNAXIP1*
rs1985437	rs8044995	1.08	SCZ	AST| HEK293T|N	1.0×10^{-3}	1.44	–	*CTRL|ELMO3|ENKD1|GFOD2|LCAT|PRMT7|RANBP10|SLC12A4|TSNAXIP1*
rs2071507	rs1042779, rs736408, rs4687552, rs2302417	1.1	ADHD|BPD|SCZ	HEK293T|N|NCL	2.6×10^{-9}	−3.36	–	*AC006254.1|GLT8D1|GLYCTK|GNL3|ITIH4|MUSTN1|NEK4|NT5DC2|POC1A|PPM1M|SFMBT1|TMEM110*
rs72708145	rs140505938	1.1	BPD|SCZ	N|NCL	5.6×10^{-3}	−0.56	*RPRD2|VPS45*	*RPRD2|SV2A|VPS45*
rs1626899	rs2102949, rs2851447	1.1	BPD|SCZ	AST|N|NCL	6.8×10^{-4}	1.27	*ABCB9|CDK2AP1|MPHOSPH9|PITPNM2*	*ABCB9|AC026362.1|ARL6IP4|CCDC62|CDK2AP1|MPHOSPH9|MTRFR|OGFOD2|PITPNM2*
rs10786713	rs7085104, rs11191424, rs11191419	1.11	BPD|SCZ} ASD	HEK293T|N|NCL	2.0×10^{-3}	−0.91	*ARL3|AS3MT|SFXN2*	*ACTR1A|ARL3|AS3MT|BORCS7|CALHM2|CNNM2|NT5C2|PFN1P11|SFXN2|SUFU|WBP1L*
rs907088	rs2517959	1.13	BPD	AST|HEK293T|N	8.5×10^{-4}	−1.12	*ERBB2|GSDMB|ORMDL3|PGAP3|PNMT*	*ERBB2|FBXL20|GSDMB|ORMDL3|PGAP3|PNMT*
rs174530	rs174576, rs28456, rs174537, rs1535, rs174535, rs4246215	1.13	BPD	N	7.4×10^{-4}	0.77	*FADS1|FADS2|TMEM258*	*FADS1|FADS2|FADS3|TMEM258*
rs903503	rs2517959	1.13	BPD	AST|N	3.6×10^{-4}	−1.08	*ERBB2|GSDMB|ORMDL3*	*ERBB2|FBXL20|GSDMB|ORMDL3|PGAP3|PNMT*
rs9266066	rs2596500	1.18	MDD|SCZ	HEK293T|N	1.3×10^{-3}	0.75	*C4A|CCHCR1|FLOT1|HLA-C|IER3|LY6G5B|MICB|MSH5|NOTCH4|RNF5|VARS2*	*C4A|CCHCR1|CYP21A1P|FLOT1|HCG22|HLA-C|IER3|LY6G5B|MICB|MSH5|NOTCH4|RNF5|VARS2*
rs174561	rs28456, rs174576	1.18	BPD	N|NCL	1.6×10^{-4}	−2.37	*FADS1|FADS3*	*FADS1|FADS2|FADS3|TMEM258*
rs762995	rs134882, rs2239612	1.22	BPD|SCZ	N	5.7×10^{-4}	−0.71	–	*CCDC134|CYP2D6|DESI1|NAGA|NDUFA6*
rs35525740	rs7914558	1.22	ADHD|BPD|SCZ	N	2.5×10^{-4}	−1.19	*ARL3|AS3MT|CALHM2|CNNM2*	*ARL3|AS3MT|CALHM2|CNNM2*
rs12264415	rs7914558	1.22	ADHD|BPD|SCZ	AST|N|NCL	5.7×10^{-4}	1.19	–	*ACTR1A|ARL3|AS3MT|BORCS7|CALHM2|CNNM2*
rs200948	rs1765142, rs34706883, rs112509803	1.24	MDD|SCZ|BPD	HEK293T|N	7.5×10^{-4}	3.07	*PRSS16|ZSCAN23|ZSCAN9*	*PGBD1|PRSS16|ZKSCAN3|ZKSCAN8|ZNF165|ZSCAN23|ZSCAN26|ZSCAN31|ZSCAN9*
rs200483	rs34706883	1.24	MDD|SCZ	N|NCL	1.8×10^{-3}	−0.53	*PGBD1|ZSCAN23|ZSCAN31|ZSCAN9*	*PGBD1|PRSS16|ZKSCAN3|ZNF165|ZSCAN23|ZSCAN26|ZSCAN31|ZSCAN9*
rs370155	rs34706883, rs45509595	1.24	MDD|SCZ|BPD	N|NCL	8.5×10^{-4}	−1.01	*PGBD1|PRSS16|ZKSCAN3|ZNF165|ZSCAN23|ZSCAN31*	*PGBD1|PRSS16|ZKSCAN3|ZNF165|ZSCAN23|ZSCAN26|ZSCAN31|ZSCAN9*
rs2428682	rs1985762	1.68	BLPD	HEK293T|N|NCL	1.5×10^{-4}	−1.09	–	*EME1*
rs3134944	rs1800625	3.78	SCZ	N	2.9×10^{-4}	0.69	–	*AGER|C4A|CYP21A1P|FKBPL|HLA-B|HLA-C|HLA-DMA|HLA-DQB1|LY6G5B|MICB|NEU1|NOTCH4|PBX2|POU5F1|RNF5|SKIV2L|ZBTB12*
rs10938176	rs34215985	1.04	MDD|BPD	AST|HEK293T	2.5×10^{-3}	−1.33	–	SLC30A9
rs301807	rs301806, rs301805, rs159963	0.03 unit desc	MDD	AST|N|NCL	3.0×10^{-8}	−1.38	–	*ENO1|RERE*

https://doi.org/10.1038/s41588-023-01533-5

Table 1 (continued) | daSNVs in neuropsychiatric diseases

daSNP	Lead SNP	OR/β	Diseases	MPRA cell type	daSNP avg *P* value	daSNP avg log_2(FC)	HiChIP genes	eGenes
rs6926869	rs6930781	0.46 unit incr	MDD	N	7.3×10^{-4}	−0.64	–	*WASF1*
rs798744	rs4818048	0.895 unit incr	OCD	HEK293T|N|NCL	7.6×10^{-4}	−1.85	*FAM53A|FGFR3|TACC3|TMEM129*	*FAM53A|FGFR3|TACC3|TMEM129*
rs2799077	rs853679	5.30 unit desc	MDD	N	1.4×10^{-3}	−0.72	*PGBD1|ZKSCAN3|ZKSCAN4|ZSCAN23|ZSCAN31*	*PGBD1|ZKSCAN3|ZKSCAN4|ZKSCAN8|ZNF165|ZSCAN23|ZSCAN26|ZSCAN31*
rs4395073	rs4785741	70.99 unit incr	OCD|ASD	N	1.9×10^{-3}	1.02	*CDK10|GAS8|SPATA2L*	*CDK10|GAS8|SPATA2L*

Selected daSNVs of the 892 significant hits. rsIDs of daSNVs and linked lead SNVs from GWAS data are listed, along with the odds ratio value reported in the GWAS Catalog. If no odds ratio is given, a ± modifier for the *β* value is listed. Diseases implicated, MPRA cell type significant, daSNV average FDR-corrected *P* value from MPRAnalyze's likelihood ratio test, average log_2(FC) (alternate/reference) and HiChIP-derived genes linked to daSNVs by chromatin looping, as well as eQTL-linked genes (eGenes), are shown. Abbreviations: N, human H9 embryonic stem cell derived neuronal cells; NCL, neuronal cell line.

each putative eGene, at an average distance of ~20 kb. Approximately 50% of GWAS gene annotations—typically annotated to the nearest gene—were discordant with eQTL or chromatin linkage (Extended Data Fig. 5), with minimal correlation between MPRA and GTEx effect sizes (Supplementary Data 3), suggesting alternative epigenetic or environmental factors affect allele-specific activity. A distilled list of chromatin data-linked genes with eQTL support (Table 1 and Supplementary Data 3) was generated to nominate disease-linked putative eGenes. For example, the 10q24.32 locus (Fig. 5) containing *CNNM2* magnesium transporter and *AS3MT* arsenic transporter genes is of interest because both have been previously associated with SCZ[40,41]. Although four daSNVs were annotated as eQTLs for the two genes in GTEx, only two daSNVs, rs12264415 and rs1046411, displayed local looping, as detected by H3K27ac HiChIP (Fig. 5b) in human neural cells. rs12264415 and rs1046411 both looped to the *CNNM2* promoter. However, no daSNVs looped to the *AS3MT* promoter. rs12264415, which displays decreased transcription-directing activity with the risk/alternate G allele versus the protective/reference T allele, is of particular interest, as MPRA-allele-specific signal was only present in neurons and not in HEK293Ts (Fig. 5e). Transcriptional dysregulation by this daSNV was predicted to be due to an AP2A motif in the risk SNV (Fig. 5c). CRISPRi of rs12264415 (Fig. 5f) decreased *CNNM2* mRNA in SH-SY5Y neural cells but not *AS3MT* or another adjacent gene, *ARL3*. Gene editing via both Cas12 and Cas9-based methods generated isogenic SH-SY5Y cells that differ only by a single nucleotide at the rs12264415 daSNV. The G disease risk SNV reduced *CNNM2* mRNA expression to 43.8% of the T SNV by Cas12-based editing. Cas9 editing produced similar results, with no significantly impacted *AS3MT* expression (Fig. 5g). The disease-linked rs12264415 daSNV modulates expression of the *CNNM2* magnesium transporter but not adjacent genes.

In humans, mutations in *CNNM2* are strongly linked to hypomagnesemia, leading to seizures and impaired brain development[40,42]. The association between magnesium and neuropsychiatric disease was therefore assessed using an 846,795-person cohort from the Department of Veteran's Affairs. An inverse relationship was observed between the prevalence of SCZ, MDD and BPD and serum magnesium levels. This was not seen for other neurologic disorders, such as Alzheimer's dementia. These trends held, when removing patients with alcohol use disorder, a possible confounder of the relationship between disease prevalence and magnesium levels. A significant difference was observed between relative disease prevalence between the bottom tenth and upper tenth deciles of serum magnesium (Extended Data Fig. 6). Integration of laboratory data with clinical diagnoses suggests that decreased magnesium correlates with an increased prevalence of specific neuropsychiatric disorders.

Noncoding regulatory risk and large effect size variants

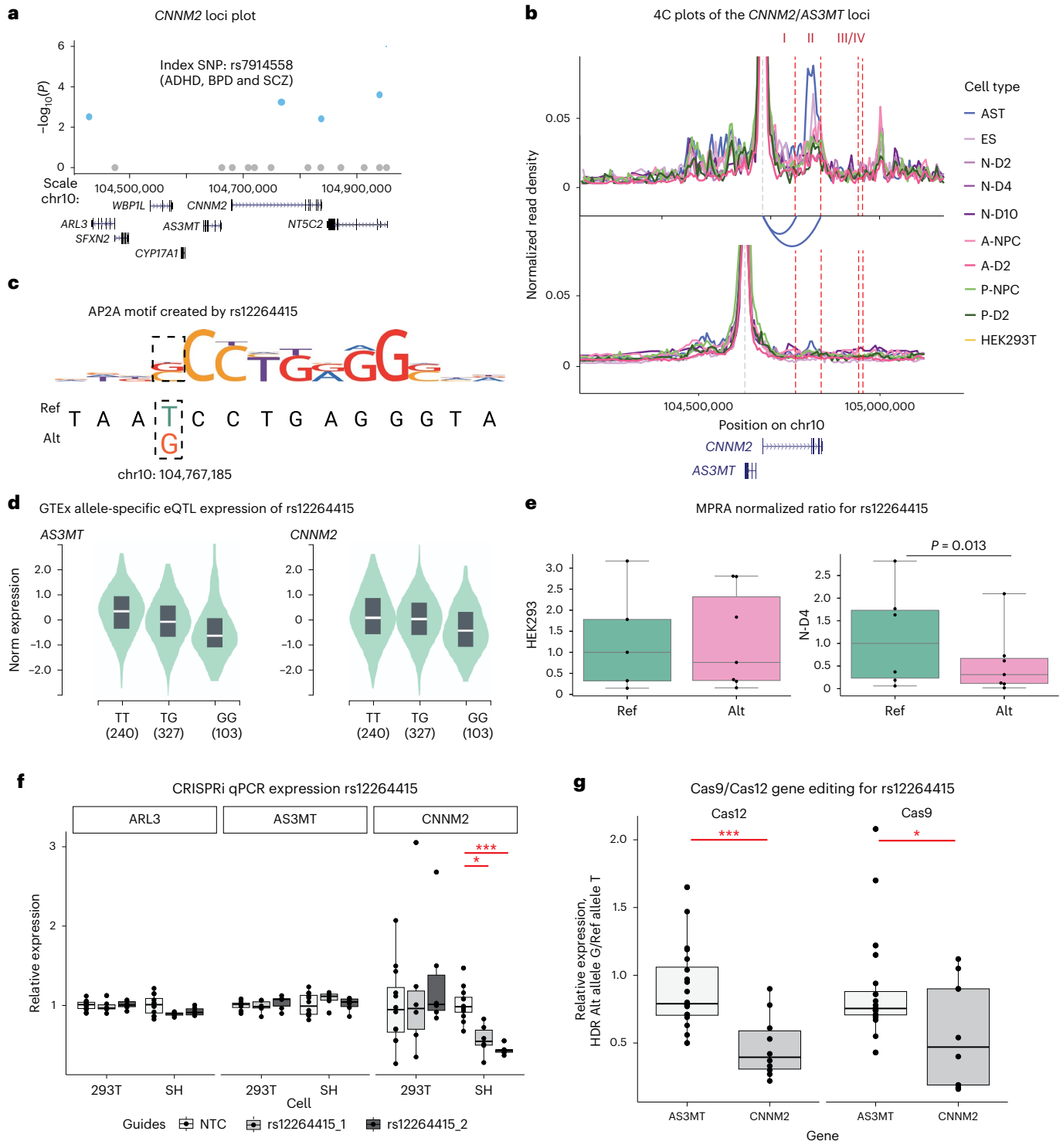
To further integrate daSNVs into the architecture of polygenic neuropsychiatric disease risk and to nominate pathomechanisms supported by orthogonal lines of evidence, analyses with larger effect size protein-coding variants were performed. The latter include coding genes whose mutation leads to Mendelian central nervous system (CNS) diseases as well as rare coding variants identified in neuropsychiatric disease risk. A total of 60 daSNV eGenes were found to be mutated in Mendelian CNS diseases (Fig. 6a,b), a significant enrichment consistent with the premise that regulatory variants produce less extreme pathologic impacts than coding mutations. For example, in the matrix of overlapping daSNV eGenes with Mendelian CNS disorder genes (Fig. 6c) is the *RERE* gene, whose deletion impairs human neurodevelopment via decreased cortical thickness[43,44]. *RERE* is a putative eGene linked to the rs301806 daSNV in MDD, suggesting a link between *RERE*-dependent

Fig. 5 | The *CNNM2* magnesium transporter gene locus. a, SNV chromosomal maps for *CNNM2* loci. Blue indicates daSNV is found to be allele-specific by MPRA, while size of the circle indicates MPRA log(FC). Index SNP rsIDs are listed. **b**, We note that rs7914558 corresponds to the daSNV rs12264415 *CNNM2* (top) and *AS3MT* (bottom) H3K27ac HiChIP 4C plots depicting looping to respective gene, color-coded by cell type. Red dashed lines indicate SNVs that are eQTLs for *CNNM2* in GTEx: I = rs12264415, II = rs1046411, III = rs35525740 and IV = rs1141095. Loops are present between SNVs I and II to *CNNM2* but not to *AS3MT*. **c**, Motif PWM analysis showing putative AP2A motif formation SNV rs12264415. **d**, GTEx allele-specific normalized expression violin plots for *AS3MT* and *CNNM2* for eQTL rs12264415 in whole blood. In the violin plots, centerline represents median, box edges represent upper and lower quartiles and distribution is derived from all relevant tissue samples on GTEx portal. **e**, Box-and-whisker plot showing normalized MPRA counts ratios for reference (teal) to alternate (pink) allele for rs12264415 in HEK293T and N-D2 tissues. Ratios are normalized to the median reference allele values, where the centerline is the median of each MPRA normalized ratio (each point is a genomic instance with at least one count, n_{max} = 5 and 6, for HEK293T Ref or Alt, respectively; and n = 6 and 7 for N-D4 Ref or Alt, respectively). FDR-corrected *P* values were calculated using MPRAAnalyze's likelihood ratio test, showing there is a significant allele-specific activity in N-D2 (P = 0.013) but not in HEK293Ts (P = 0.88). **f**, CRISPRi box-and-whisker plot showing relative qPCR expression of *ARL3*, *AS3MT* and *CNNM2* at rs12264415 loci in both HEK293s and SH-SY5Y cells. *P* values are calculated from two-sided Student's *t* tests with $^{*}P = 8.6\times10^{-4}$ and $^{***}P = 4.0\times10^{-7}$ for n = 6 biological replicates. **g**, Box-and-whisker plots showing relative Cas12 (left) and Cas9 (right)-based gene editing expression of allele G (alternate) versus allele T (reference) at the rs12264415 loci. Two-sided Student's *t* test, $^{*}P = 0.036$ and $^{***}P = 6.9\times10^{-4}$; n = 5 biological replicates. All box-and-whisker subplots in this figure are shown with a maximum whisker length of 1.5× IQR. The centerline represents the median and the box edges represent the upper and lower quartiles. All outliers are shown.

https://doi.org/10.1038/s41588-023-01533-5

neurodevelopmental processes and risk for major depression. In this regard, rs301806 disrupted the DNA binding motif for *RUNX1*, a TF with essential roles in neural differentiation[45] and also decreased MPRA signal in early neuronal differentiation. Episomal ChIP PCR for RUNX1 demonstrated differential binding of RUNX1 for rs301806 (Extended Data Fig. 7). This MDD-linked daSNV may thus impact neurodevelopmental pathways by altering TF binding and subsequent expression of *RERE*, supporting the rationale for systemic studies of daSNVs' action as eQTLs for genes whose coding mutations produce Mendelian disorders of the CNS.

The overlap between daSNV eGenes and rare coding variants identified by neuropsychiatric disease GWAS was next explored. Genes found mutated in large-scale exome sequencing from the Schizophrenia Exome Sequencing Meta-analysis (SCHEMA) consortium[46] were intersected with genes linked to daSNVs via neural cell chromatin architecture ($n = 7$ genes; *C4A*, *CACNA1G*, *DAGLA*, *MAGI2*, *STAF1*, *SV2A* and *XPO7*; Fig. 6d–g and Supplementary Table 10). For example, promoters for protein-coding variant genes identified in SCZ exome sequencing, including *CACNA1G* and *DAGLA*, were in physical contact



with rs2428682 and rs174568 daSNVs, respectively (Fig. 6e–h). Given that SCHEMA-prioritized protein variants of *CACNA1G* contribute to SCZ risk—potentially by dysregulating calcium influx during neural action potentials, which in turn may disrupt neurotransmitter release and neural circuitry important in disease risk[47]—noncoding variants with small effect sizes altering *CACNA1G* expression may exert similar impact. These findings are consistent with the premise that regulatory daSNVs may pathogenically alter the expression of the same genes whose coding mutations exert large effect size impacts on neuropsychiatric disorders.

Determining shared symptom manifestations of regulatory risk

Neuropsychiatric diseases are diagnosed based on constellations of clinical symptoms. Many symptoms are shared among disorders, suggesting they may reflect common genetic bases. Consistent with this, 192 of 641 daSNV-linked candidate eGenes (30.0%) were shared between at least two disorders. BPD and BPLD had the highest fraction of their eGenes shared with at least one other disease, suggesting they arise via dysregulated biologic processes impacted in multiple conditions (Fig. 7a). eGenes were associated with 806 psychiatric disease symptom codes derived from UK Biobank GWAS summary statistics[48], clustered into 64 annotated phenotypes based on gene profile, and normalized for gene–symptom relationships (Fig. 7b). To avoid over-annotation of individual daSNVs, loci containing multiple disease genes were collapsed to show only one gene per locus, based on prioritization schema favoring genes with brain-specific expression modulated by, chromatin linkage to, and prior literature evidence for, the given daSNV. Anxiety-relevant symptoms such as worry, tenseness and stress were enriched in inflammatory daSNV eGenes around the *C4A* and *HLA-C* locus. Interestingly, CYP2D6, a cytochrome P450 enzyme with a compound metabolizing role in CNS, displayed protective gene-trait effects for anxiety-related symptoms, indicating the potential for therapeutic targeting. Further supporting this, known therapeutics were linked to putative eGenes via connectivity map (CMAP)[49] perturbation analysis (Extended Data Fig. 8), highlighting eight daSNV-linked genes, including *SV2A, AP3B2* and *ARC*, as higher priority drug targets (see Supplementary Table 11 for the full list of prioritized drug targets). *GNL3* and *POC1A* genes were linked to another clinical feature, psychosis history. Together, these gene-phenotype associations identify potential shared molecular bases for individual neuropsychiatric symptoms.

Gene-phenotype associations were further examined for clues to molecular etiology of shared pathomechanisms underlying common neuropsychiatric symptoms. Loci maps were created at known MPRA daSNV 'hotspots' *CYP2D6*, *TOR1A*, *GNL3*, *HLA-C* and *GIT1* (Fig. 7c–e), to determine the interplay between daSNVs associated with multiple disorders. The following three general structures were observed: one where a single-GWAS-lead SNV was associated with multiple disorders (*TOR1A*), another where multiple disorders shared multiple SNVs at a given locus (*GIT1*), and most commonly, where individual diseases have different sets of associated SNVs (*CYP2D2*, *HLA-C*, *CNNM2*, *GNL3* and *RERE*) that modulate the same set of genes. This pattern was further defined by generating daSNV–gene networks surrounding shared clinical features and biological processes, such as psychosis history (Fig. 7f), synaptic signaling (Fig. 7g), regulation of cytokine production, sleep issues, anhedonia and irritability (Supplementary Data 4). For psychosis history, pertinent for SCZ and BPD, the latter two diseases shared daSNVs at loci surrounding the gene for *GNL3*, a brain-expressed G protein important for stem cell proliferation and differentiation[50], and for *KCTD10*, a brain-expressed protein-mediating tetramerization of voltage-gated potassium channel subunits[51]. Both loci may have putative linkages to psychosis in multiple disorders[52,53]. SCZ also displayed two additional daSNVs relevant for psychoses, namely rs746011 and rs3824756. Similar observations were made in synaptic signaling, which was linked to daSNVs in four different diseases. These data suggest that common genetic pathomechanisms mediate symptomology shared across multiple neuropsychiatric diseases.

Discussion

We explored the genetic contributions of noncoding SNVs to neuropsychiatric disease by integrating MPRA with epigenomic profiling in human neural stem cells, differentiating human neurons, neural cell lines and primary AST, and thereby generated a multidisorder, single-nucleotide compendium of 892 functional daSNVs in regulatory DNA linked to risk for ten neuropsychiatric diseases. Altered transcription-directing activity of these daSNVs was not associated with major impacts on chromatin accessibility and looping, suggesting action via other mechanisms, such as modifying TF binding, as observed for RUNX1 at rs301806. Network analyses linking daSNVs to putative eGenes demonstrated that disease-centric gene networks recurrently implicate several biologic processes in neuropsychiatric disease, namely ion channel disruption (particularly, calcium, potassium, chloride and magnesium channels), modulation of neuronal proliferation (including genes such as *NEK4* and *GNL3*), protocadherin function, Notch signaling, synaptic action via *SV2A* and immune activity (particularly MHC and C4 complement families). Chromatin and expression linkages suggested that neuropsychiatric disease pathogenesis unfolds primarily in neuronal cells, highlighted disease-linked eGenes, such as *POU5F1, CNNM2* and *RERE*, and mapped putative disease eGenes to native brain cortical subpopulations.

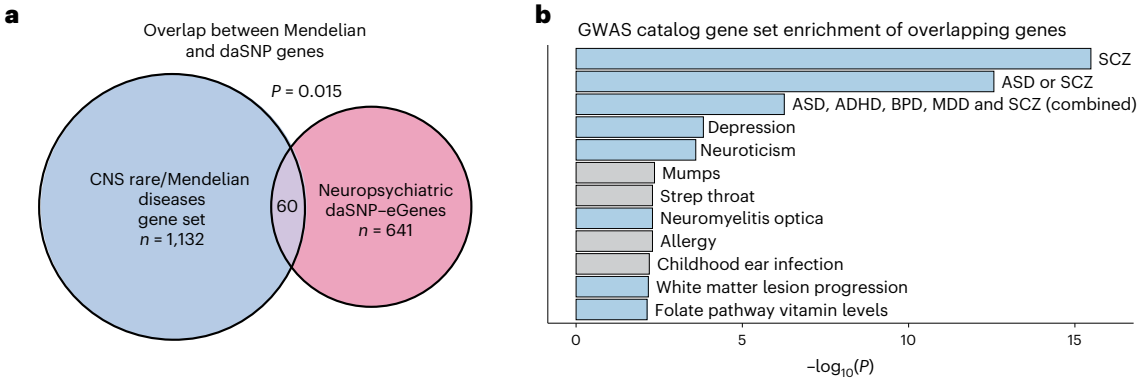
CRISPRi and gene editing demonstrated that specific neuropsychiatric risk-linked daSNVs modulate the expression of *CNNM2* gene. The CNNM2 transporter is important in magnesium homeostasis. Consistent with a role for magnesium in neuropsychiatric disorders are its known roles in neuronal development, transmission, and survival as are reports of its use as an off-label supplement for depression[54].

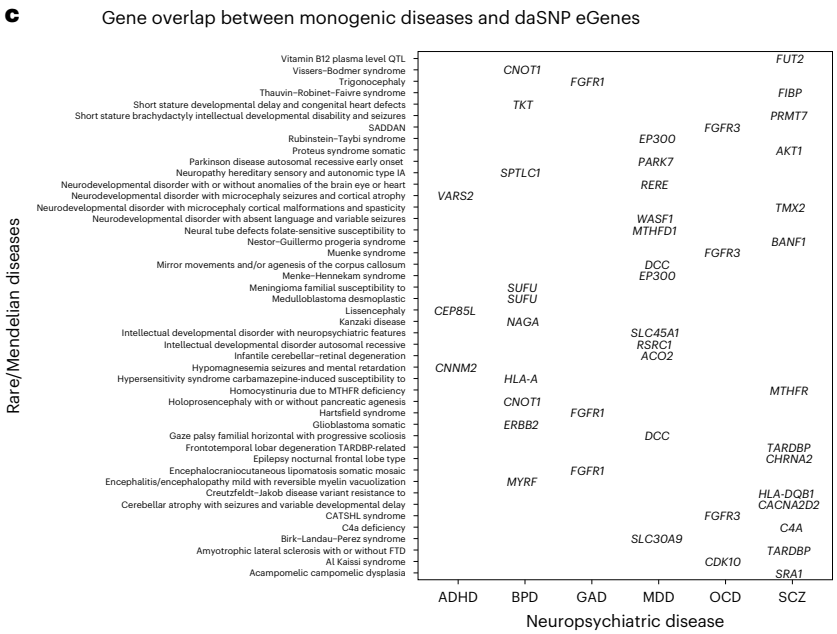
Fig. 6 | Altered coding genes in CNS diseases inform risk in psychiatric disorders. a, Venn diagram depicting overlap between Mendelian CNS disease genes and the neuropsychiatric eGenes linked to daSNVs; *P* value = hypergeometric test between the two gene sets over a background of all potential disease-associated genes ($n = 15{,}999$ possible Mendelian genes). **b**, Gene set enrichment analysis calculated by EnrichR with Benjamin–Hochberg-corrected *P* value from a two-sided hypergeometric test for the 2019 GWAS Catalog of the 60 overlapping genes, where the blue bars indicate diseases of neuropsychiatric etiology or linkage. **c**, Grid chart of genes in the intersection between rare/Mendelian diseases and neuropsychiatric diseases. **d**, Abbreviated table of SCZ coding variant genes linked to chromatin data (see also Supplementary Table 1). **e**, Tracks for *CACNA1G*, where the peak tracks show the log(FC) change from cell-type-specific MPRA for the daSNVs, and the bottom loop track shows the looping data for P-NPC cell type, indicating the daSNV rs2428682 loops to the promoter of *CACNA1G*. Scales are only included if there was a peak within the given region shown. **f**, Box-and-whisker plot showing normalized MPRA counts ratios for reference (teal) to alternate (pink) allele for rs2428682 in P-NPC, where the centerline is the median of each MPRA normalized ratio (each point is a genomic barcode instance with at least one count, $n = 7$ for Ref, $n = 10$ for Alt), box limits are the upper and lower quartiles, whiskers are the 1.5× IQR, and points shown are outliers. Ratios are normalized to the median reference allele values. FDR-corrected *P* values were calculated using MPRAnalyze's likelihood ratio test to indicate significant allele-specific activity ($P = 3.3 \times 10^{-5}$). **g**, Similar track for coding variant gene *DAGLA*. **h**, Box-and-whisker plot showing normalized MPRA ratios for one of the daSNVs linked to DAGLA, rs174568, where each point is a barcode ($n = 8$ for Ref, $n = 9$ for Alt). All boxplots shown have a maximum whisker length of 1.5× IQR. The centerline represents the median; the box edges represent the upper and lower quartiles. All outliers are shown. FDR-corrected *P* values were calculated using MPRAnalyze's likelihood ratio test to indicate significant allele-specific activity ($P = 0.034$).

https://doi.org/10.1038/s41588-023-01533-5

The relative impacts of altered magnesium levels and transport in the brain versus total magnesium body stores and the potential impact of magnesium supplementation on the prevention and treatment of neuropsychiatric disease are of interest for further study.

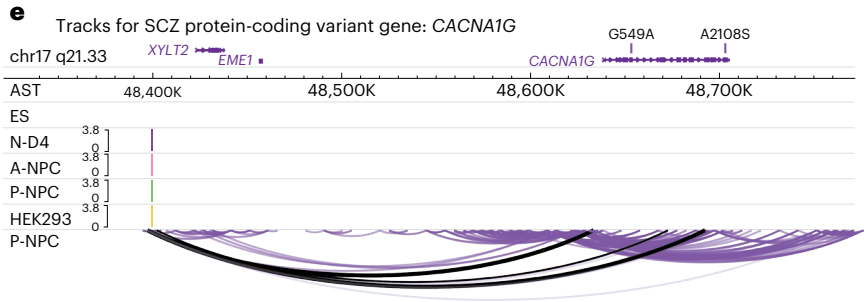
MPRA during neuronal differentiation helped identify functional daSNVs that may contribute to an interplay between neurodevelopmental milestones and neuropsychiatric disease risk—as observed here in early neuronal differentiation for the MDD-linked *RERE* eGene.

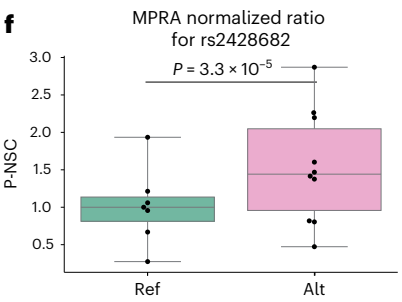


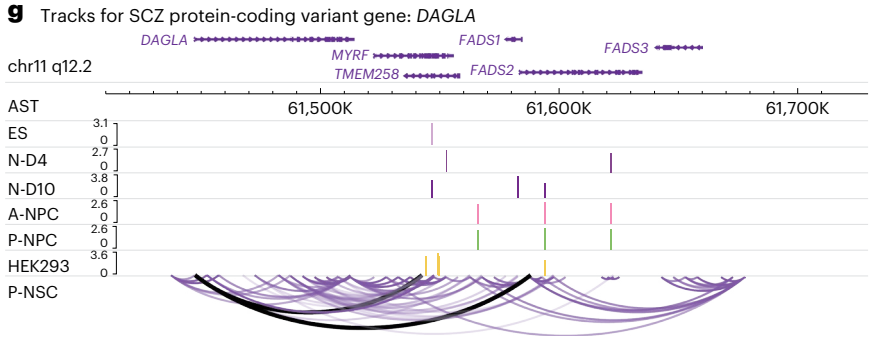


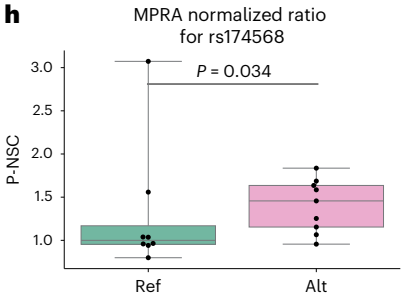
d SCZ coding variant genes linked to daSNPs

Gene	daSNPs	Missense mutations + PTVs
C4A	rs9265985, rs2854002, rs9266066, rs2442722, rs813115, rs3096696, rs3130347, rs3134944, rs3134940, rs2071278,	Note 1
CACNA1G	rs2428682	A2108S, Gly529A, c.6060+2T>C, S818AfsTer21, W925Ter, Q968Ter, Leu1050HfsTer38, W1488Ter, L1685RfsTer27, c.5227-2A>G, c.5925+1G>T
DAGLA	rs174529, rs174530, rs174533, rs174534, rs174537, rs174541, rs174561, rs174568, rs174599	L401VfsTer8, c.1213-2A>G, Q451Ter, Y497Ter, c.1514+1G>T, R547Ter, c.2171+1G>A, A843CfsTer158, A1032GfsTer5
MAGI2	rs322004	F163LfsTer11 , Q1305RfsTer169, R1084Ter, P908LfsTer32, c.2311+2C>A, E531Ter, I473SfsTer4, c.1225+1G>A, E238Ter, W11Ter
STAG1	rs900947	E1235Ter, R1206Ter, L927YfsTer15, I839MfsTer56, R131Ter, R51Ter
SV2A	rs72708145	R390Ter, F718CfsTer17, R507Ter, V486SfsTer13, Q476Ter, E138GfsTer15, R67Ter, R43Ter
XPO7	rs746011, rs11136093, rs11780207	Note 2

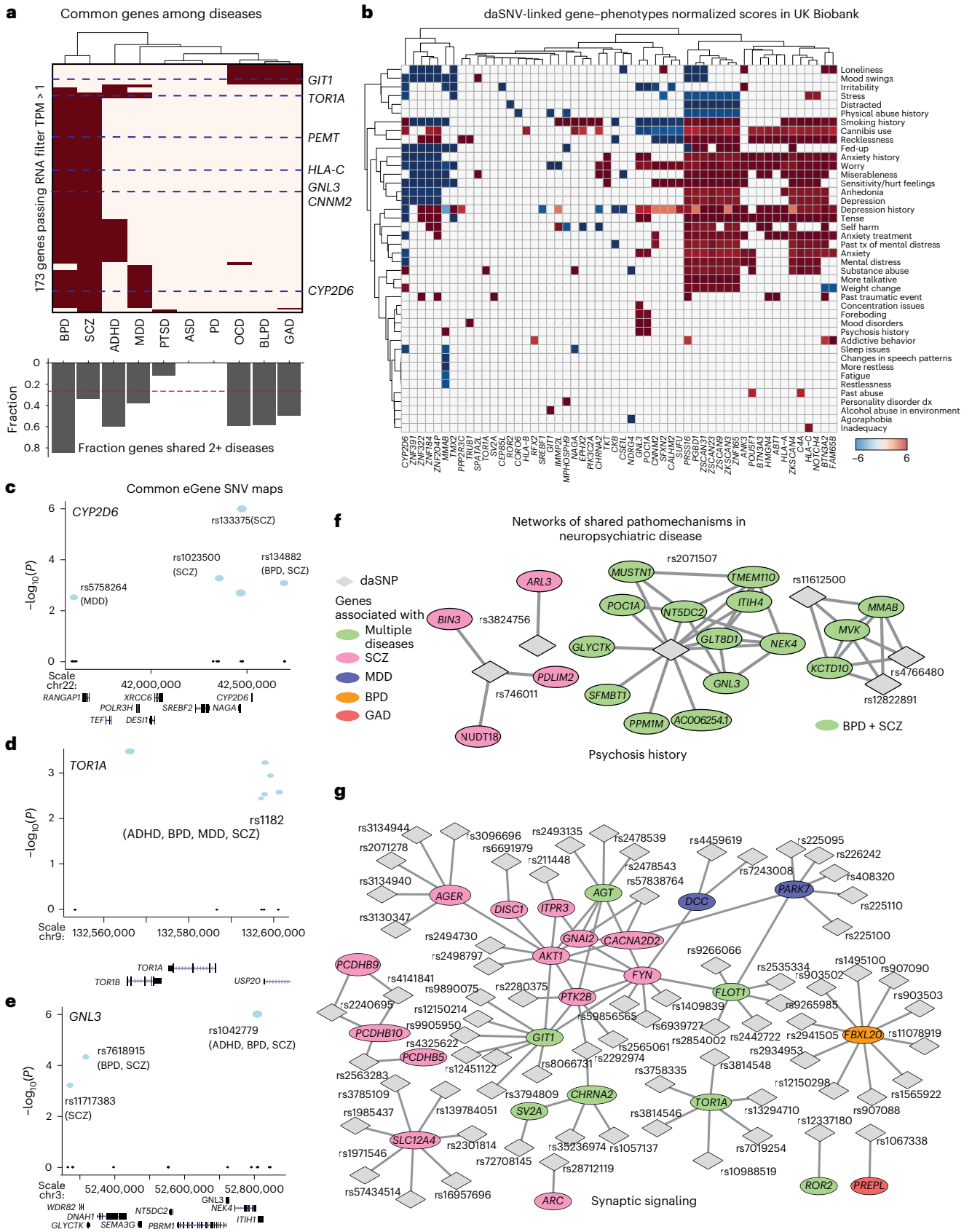








https://doi.org/10.1038/s41588-023-01533-5



Tissue-specific diseases have cell-type-specific pathomechanisms. However, prior large-scale MPRA studies of human regulatory variants were conducted in transformed cell lines, such as HEK293T and K562 cells, that may not reflect a disease-relevant gene regulatory milieu. Consistent with this, HEK293T cells failed to detect the majority of functional daSNV alleles in this study. Overlap between SNVs studied here and other datasets[19–21,55] was low; however, rs214469 was studied in ref. 55 with concordant directionality in allelic activity in the neural cell

Fig. 7 | daSNV–eGene symptom linkage in neuropsychiatric disorders. **a**, eGene (row) by disease binary heatmap where red indicates 1 of 173 putative daSNV eGenes associated with at least two diseases and expressed in the corresponding cell type with TPM > 1. Heatmap clusters BPD and SCZ eGenes as being the most similar. Below the heatmap, a bar chart displaying fraction of eGenes shared between two or more diseases is shown. The red dashed line indicates that 26.8% (173 of 641 eGenes) overall are shared between two or more diseases. **b**, UK Biobank PheWAS analysis shown as a heatmap of mean normalized β values for UK Biobank neuropsychiatric symptoms across conditions for the different eGenes. eGenes are clustered by chromosomal location. **c**–**e**, daSNV chromosomal maps for common eGenes—*CYP2D6* (**c**), *TOR1A* (**d**) and *GNL3* (**e**)—implicated in multiple diseases with ≥1 index SNVs directed from literature or GWAS. Blue indicates daSNV found to be allele-specific by MPRA, while the size of the circle indicates the absolute value of the MPRA $\log_2$(FC) (alternate/reference). **f**, daSNV (diamond)–gene (ovals) networks of shared pathomechanisms in neuropsychiatric disease. Network of daSNV–eGene candidates implicated in psychosis history was derived from UK Biobank, where green ellipses are genes shared between BPD and SCZ, while pink nodes are SCZ only. **g**, The network of daSNV–eGene candidates implicated in the GO biological process synaptic signaling. Genes are color-coded by disease of origin, where the green circles represent implicated genes shared between multiple diseases. Genes are linked via StringDB v11. Tx, treatment; dx, diagnosis.

type (SH-SY5Y) and opposing directionality in the non-neural cell type (K562), highlighting the issue of cell type context for MPRA. Cellular diversity in the brain is enormous; however, because ES-derived neurons are more fetal-like rather than adult-like, some daSNVs active in mature neurons, such as those mediating synaptic transmission, may have been missed in the current study. Future work may extend efforts to include functional studies in additional brain cell types as well as cells from other tissues that contribute to pathogenesis, notably the immune system.

Accurately assigning target genes to noncoding variants within intact developing human tissue remains challenging. Neuropsychiatric diseases can unfold over decades, with environmental modulation of the activity of regulatory variants, which may account for the lack of a strong correlation between MPRA and eQTL signals. Amalgamating multiple orthogonal approaches—not only tissue-specific eQTL–gene mapping[13] but also network-based analyses, literature annotation, cell-type-matched chromatin looping, functional assays in relevant cell types and new predictive models—may help prioritize targets for further study. We observed that putative daSNV eGenes were enriched in genes mutated in Mendelian CNS diseases and coding variants identified in SCZ exome sequencing, indicating that these genes may be altered in multiple ways to promote the pathogenesis of neuropsychiatric disease. The link from variant, to gene, and gene function in the appropriate tissue to clinical outcomes remains a major goal in polygenic disease.

Cross-disorder analysis of neuropsychiatric genetic risk may help uncover the genetic basis of common psychiatric clinical symptoms and translate variant–gene linkage information into useful personalized therapies. The roughly 22% of eGenes identified here, which are common between multiple disorders, nominate shared biologic processes, primarily those impacting neural development. Gene–disease connections were extended to nominate gene–symptom linkages, where symptoms are clinical features, such as anxiety or psychoses, shared across multiple neuropsychiatric disorders. Such a gene–symptom approach to disease classification highlights well-known difficulties in establishing psychiatric diagnoses for many persons whose combination of clinical features fail to fit into current diagnostic schemas. Extending such gene–symptom linkage may advance efforts toward more precise neuropsychiatric disease subtyping and symptom-targeted therapies. Regarding the latter, identification of pathogenic risk networks shared across multiple neuropsychiatric disorders may help apply pharmacogenomic understanding to commonly impacted biological processes. For example, inherited mutations in *TOR1A*, an eGene implicated above in multiple neuropsychiatric diseases, cause dystonia[56], a known side effect of some antidepressants and antipsychotics. Defining associations between *TOR1A* daSNVs and dystonia-related medication side effect profiles may assist personalized approaches to psychiatric drug treatment in multiple diseases. Enhanced understanding of regulatory risk variants may thus have the potential to improve both molecular classification and treatment of human neuropsychiatric disorders.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01533-5.

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
2. PsychENCODE Consortium et al.The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
3. Ombrato, L. et al. Metastatic-niche labelling reveals parenchymal cells with stem features. *Nature* **572**, 603–608 (2019).
4. Witt, S. H. et al. Genome-wide association study of borderline personality disorder reveals genetic overlap with bipolar disorder, major depression and schizophrenia. *Transl. Psychiatry* **7**, e1155 (2017).
5. Meier, S. M. et al. Genetic variants associated with anxiety and stress-related disorders: a genome-wide association study and mouse-model study. *JAMA Psychiatry* **76**, 924–932 (2019).
6. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
7. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).
8. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
9. Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
10. Mulvey, B., Lagunas, T. & Dougherty, J. D. Massively parallel reporter assays: defining functional psychiatric genetic variants across biological contexts. *Biol. Psychiatry* **89**, 76–89 (2021).
11. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
12. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
13. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
14. Zhang, Y. et al. Rapid single-step induction of functional neurons from human pluripotent stem cells. *Neuron* **78**, 785–798 (2013).
15. Ashuach, T. et al. MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
16. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
17. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC–seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–9 (2015).
18. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

19. Song, M. et al. Mapping *cis*-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019).
20. Zhang, S. et al. Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* **369**, 561–565 (2020).
21. Inoue, F. et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity.*Genome Res.* **27**, 38–52 (2017).
22. Herdegen, T. & Waetzig, V. AP-1 proteins in the adult brain: facts and fiction about effectors of neuroprotection and neurodegeneration. *Oncogene* **20**, 2424–2437 (2001).
23. Chew, L. J. et al. Sox17 regulates a program of oligodendrocyte progenitor cell expansion and differentiation during development and repair. *Cell Rep.* **29**, 3173–3186 (2019).
24. Ruiz i Altaba, A., Palma, V. & Dahmane, N. Hedgehog–GLI signaling and the growth of the brain. *Nat. Rev. Neurosci.* **3**, 24–33 (2002).
25. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
26. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
27. Hook, P. W. & McCallion, A. S. Leveraging mouse chromatin data for heritability enrichment informs common disease architecture and reveals cortical layer contributions to schizophrenia. *Genome Res.* **30**, 528–539 (2020).
28. Bulik-Sullivan, B. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
29. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
30. Kinney, D. K. et al. A unifying hypothesis of schizophrenia: abnormal immune system development may help explain roles of prenatal hazards, post-pubertal onset, stress, genes, climate, infections, and brain dysfunction. *Med. Hypotheses* **74**, 555–563 (2010).
31. Shao, Z. et al. Dysregulated protocadherin-pathway activity as an intrinsic defect in induced pluripotent stem cell-derived cortical interneurons from subjects with schizophrenia. *Nat. Neurosci.* **22**, 229–242 (2019).
32. Hoseth, E. Z. et al. Attenuated Notch signaling in schizophrenia and bipolar disorder. *Sci. Rep.* **8**, 5349 (2018).
33. Onwordi, E. C. et al. Synaptic density marker SV2A is reduced in schizophrenia patients and unaffected by antipsychotics in rats. *Nat. Commun.* **11**, 246 (2020).
34. Yabut, O. R., Fernandez, G., Huynh, T., Yoon, K. & Pleasure, S. J. Suppressor of fused is critical for maintenance of neuronal progenitor identity during corticogenesis. *Cell Rep.* **12**, 2021–2034 (2015).
35. Nord, A. S., Pattabiraman, K., Visel, A. & Rubenstein, J. L. R. Genomic perspectives of transcriptional regulation in forebrain development. *Neuron* **85**, 27–47 (2015).
36. Zhang, S. et al. OCT4 and PAX6 determine the dual function of SOX2 in human ESCs as a key pluripotent or neural factor. *Stem Cell Res. Ther.* **10**, 122 (2019).
37. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
38. Nichol, H., Amilhon, B., Manseau, F., Badrinarayanan, S. & Williams, S. Electrophysiological and morphological characterization of Chrna2 cells in the subiculum and Ca1 of the hippocampus: an optogenetic investigation. *Front. Cell. Neurosci.* **12**, 32 (2018).
39. Giralt, A. et al. Pyk2 modulates hippocampal excitatory synapses and contributes to cognitive deficits in a Huntington's disease model. *Nat. Commun.* **8**, 15592 (2017).
40. Arjona, F. J. et al. CNNM2 mutations cause impaired brain development and seizures in patients with hypomagnesemia. *PLoS Genet.* **10**, 1004267 (2014).
41. Li, M. et al. A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nat. Med.* **22**, 649–656 (2016).
42. Franken, G. A. C. et al. Cyclin M2 (CNNM2) knockout mice show mild hypomagnesaemia and developmental defects. *Sci. Rep.* **11**, 8217 (2021).
43. Adam, M. P., Mirzaa, G. M., Pagon, R. A., Wallace, S.E., Bean, L. J. H., Gripp, K. W. & Amemiya, A. (eds.). *GeneReviews* (University of Washington, 1993).
44. Katsuki, A. et al. A single-nucleotide polymorphism influences brain morphology in drug-näve patients with major depressive disorder. *Neuropsychiatr. Dis. Treat.* **15**, 2425–2432 (2019).
45. Logan, T. T., Rusnak, M. & Symes, A. J. Runx1 promotes proliferation and neuronal differentiation in adult mouse neurosphere cultures. *Stem Cell Res.* **15**, 554–564 (2015).
46. Singh, T., Neale, B. M., Daly, M. J. et al. Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.18.20192815 (2022).
47. Andrade, A. et al. Genetic associations between voltage-gated calcium channels and psychiatric disorders. *Int. J. Mol. Sci.* **20**, 3537 (2019).
48. Davis, K. A. S. et al. Mental health in UK Biobank: development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. *BJPsych Open* **6**, e18 (2020).
49. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
50. Yaghoobi, M. M., Mowla, S. J. & Tiraihi, T. Nucleostemin, a coordinator of self-renewal, is expressed in rat marrow stromal cells and turns off after induction of neural differentiation. *Neurosci. Lett.* **390**, 81–86 (2005).
51. Teng, X. et al. KCTD: a new gene family involved in neurodevelopmental and neuropsychiatric disorders. *CNS Neurosci. Ther.* **25**, 887–902 (2019).
52. Imbrici, P., Camerino, D. C. & Tricarico, D. Major channels involved in neuropsychiatric disorders and therapeutic perspectives. *Front. Genet.* **4**, 76 (2013).
53. Meng, Q. et al. Integrative analyses prioritize GNL3 as a risk gene for bipolar disorder. *Mol. Psychiatry* **25**, 2672–2684 (2020).
54. Eby, G. A. & Eby, K. L. Rapid recovery from major depression using magnesium treatment. *Med. Hypotheses* **67**, 362–370 (2006).
55. Myint, L. et al. Testing the regulatory consequences of 1,049 schizophrenia associated variants with a massively parallel reporter assay. Preprint at *bioRxiv* https://doi.org/10.1101/447557 (2019).
56. Calakos, N. et al. Functional evidence implicating a novel TOR1A mutation in idiopathic, late-onset focal dystonia. *J. Med. Genet.* **47**, 646–650 (2010).
88. Duncan L, Yilmaz Z, Gaspar H, et al. Significant Locus and Metabolic Genetic Correlations Revealed in Genome-Wide Association Study of Anorexia Nervosa. *Am. J. Psychiatry* **174**, 850–858 (2017).
89. Demontis, D., Walters, R.K., Martin, J. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
90. Grove, J., Ripke, S., Als, T.D. et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

91. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address: douglas.ruderfer@vanderbilt.edu; Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705–1715 (2018).
92. Stahl, E.A., Breen, G., Forstner, A.J. et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet*. **51**, 793–803 (2019).
93. Pardiñas, A.F., Holmans, P., Pocklington, A.J. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet*. **50**, 381–389 (2018).
94. Wray, N.R., Ripke, S., Mattheisen, M. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet*. **50**, 668–681 (2018).
95. Howard, D.M., Adams, M.J., Clarke, TK. et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci*. **22**, 343–352 (2019).
96. Nagel, M., Jansen, P.R., Stringer, S. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet*. **50**, 920–927 (2018).
97. International League Against Epilepsy Consortium on Complex Epilepsies. Electronic address: epilepsy-austin@unimelb.edu.au. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol*. **13**, 893–903 (2014).
98. Jansen, I.E., Savage, J.E., Watanabe, K. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet*. **51**, 404–413 (2019).
99. Marioni, R.E., Harris, S.E., Zhang, Q. et al. GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* **8**, 99 (2018).
100. Corces, M.R., Shcherbina, A., Kundu, S. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat Genet* **52**, 1158–1168 (2020).
101. Ferreira MAR, Mathur R, Vonk JM, et al. Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. *Am J Hum Genet*. **104**, 665–684 (2019).
102. de Lange, K., Moutsianas, L., Lee, J. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256–261 (2017).

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

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023

[1]Stanford Program in Biomedical Informatics, Stanford University, Stanford, CA, USA. [2]Program in Epithelial Biology, Stanford University, Stanford, CA, USA. [3]Department of Pathology, Stanford University, Stanford, CA, USA. [4]Department of Bioengineering, Stanford University, Stanford, CA, USA. [5]Institute for Stem Cell Biology & Regenerative Medicine, Stanford University, Stanford, CA, USA. [6]Institute of Neurobiology, Xi'an Jiaotong University Health Science Center, Xi'an, China. [7]Department of Genetics, Stanford University, Stanford, CA, USA. [8]Stanford Program in Cancer Biology, Stanford University, Stanford, CA, USA. [9]Department of Biology, Stanford University, Stanford, CA, USA. [10]Bitscopic Inc., Los Angeles, CA, USA. [11]Public Health Surveillance and Research, Department of Veterans Affairs, Washington, DC, USA. [12]Division of Infectious Disease & Geographic Medicine, Stanford University School of Medicine, Stanford, CA, USA. [13]Program in Neurobehavioral Genetics, Semel Institute, UCLA, Los Angeles, CA, USA. [14]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. [15]Veterans Affairs Palo Alto Healthcare System, Palo Alto, CA, USA.
✉e-mail: khavari@stanford.edu

Methods

Statistics and reproducibility

Sample sizes for each experiment and statistical power calculations are noted in each of the analyses sections in Methods. Sample sizes were chosen to provide sufficient confidence to validate methodological conclusions. Specifically for barcode number, power analysis for different levels of barcodes at four different $\log_2$(fold change) thresholds was performed (Extended Data Fig. 1d). The sample size of 2,221 variants was chosen based on technical feasibilities during cloning and library preparation that limited the final library size. Replicates are noted with respect to each experiment performed. All experiments were performed with two or four biological replicates and technical duplicates. MPRA data for H9 day 28, HMC3 microglia and NT2 human embryonal carcinoma cell line were excluded from the final analysis due to poor replicability and sequencing quality. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

Statistical analyses were performed with R version 3.6.1 and Python 3.7.4 in Jupyter Notebook. Parameters such as the number of replicates, the number of independent experiments, measures of center, dispersion and precision (mean ± s.d. or s.e.m.), statistical test and significance are reported in figures and figure legends, such as in Figs. 5 and 6, Extended Data Figs. 4 and 7. Raw sequencing data were processed on the Stanford Sherlock cluster. All boxplots shown have a maximum whisker length of 1.5× IQR. The centerline represents the median, the box edges represent the upper and lower quartiles, and any points visible are outliers. Data distribution was assumed to be normal but this was not formally tested. Wherever possible, nonparametric tests (such as *U* tests) were used to avoid assumptions of normality.

Analyses described in the subsections below are central to the generation of the main figures 1–7. Additional information can be found in Supplementary Methods.

Ethics statement

This research complies with all relevant ethical regulations. The use of human ES cells (H9) was carried out in accordance with Stanford University and its Center for Human Embryonic Stem Cell Research and Education. The serum magnesium study in the VA cohort project was approved by the Stanford University Institutional Review Board under the protocol entitled 'Public Health Surveillance in the Department of Veterans Affairs'. As the project was considered minimal risk, consent to participate was not required. Bitscopic operates under a 10-year Research and Development Agreement with the VA signed in 2019.

Materials availability

Two new plasmids are generated and used, pGreenFire_blastocidin (Addgene, 174103) and pD2_miniluc_newP7 (Addgene, 174105). Materials will be available, subject to materials transfer agreements. All primer sequences are included in the supplementary tables (Table S15). Please contact the authors for further information.

MPRA library

MPRA library design and variant selection. Variants for the ten neuropsychiatric conditions were selected by collating SNVs listed in GWAS Catalog[57] (December 2018) and curated psychiatric GWAS papers for a total of 13,956 index SNVs. Additionally, ClinVAR[58] SNVs linked to putative psychiatric-linked genes, nominated by PsychENCODE[2] (http://resource.psychencode.org/) and OMIM[59] (https://www.omim.org/), were added (n = 1,948 SNVs). Linked SNVs were determined by using LD information from Haploreg v4 (ref. 60), filtering for $r^2 > 0.8$ for a total of 268,545 SNVs. Additionally, SNVs were filtered through ENCODE DHS[61] immune and neuron cell narrow peaks, as listed in Supplementary Data 5 (DHS tissue filter for MPRA) as well as associated with eQTLs from PsychENCODE[2] and GTEx v7 (ref. 13), yielding 61,134 SNVs remaining.

Given the constraints of the final MPRA library size, SNVs included in the final MPRA library were prioritized based on annotated information. If a GWAS P value for the SNV was available, only SNVs with $P < 1 \times 10^{-6}$ were kept. For SCZ, BPD and ADHD, a more stringent P value threshold of 1×10^{-8} was used. Additionally, SNVs were annotated using CADD v1.4 (ref. 62). SNVs with predicted motifs and CADD phred-like c scores >20 were included, if possible. Additionally, we note that we selected signal-negative controls that are in blacklisted regions by ENCODE hg19 (n = 22) to test via MPRA.

We designed a set of DNA fragments by obtaining the genomic sequence corresponding to a 145 bp window centered at each variant of interest (hg37). For each variant, a reference and alternative sequence was designed, corresponding to each allele. In cases where multiple alternate alleles were given for an SNV, all sequence possibilities were included. Sequences were filtered to ensure the restriction sites for EcoRI, BamHI, XhoI and XbaI were not present. For XbaI, sequences were additionally filtered such that dam methylation recognition sequences would not be present. This process yielded 2,221 SNV locations to be tested.

Each MPRA library oligo included the following, in order: an MPRA forward primer, the 145 bp genomic instance sequence, a XhoI restriction site, a 10 bp randomly generated filler sequence, a XbaI restriction site, a 20 bp barcode and an MPRA reverse primer. The 20 bp barcodes are all a minimum Hamming distance of 3 apart. Each unique genomic instance is barcoded ten times. This yields a 44,400–230 bp oligo library that was synthesized by the Agilent HiFi synthesis process. Library cloning is described in Supplementary Methods.

Cell culture. Infection and culture of normal human AST (Lonza, CC-2565), HEK293T (Takara, 632180), SH-SY5Y neuroblastoma cells (ATCC, CRL-2266), IMR-32 neuroblastoma cells (ATCC, CCL-127), D283 medulloblastoma cells (ATCC, HTB-185) and D341 medulloblastoma cells (ATCC, HTB-187) for the MPRA experiments are described in Supplementary Methods.

Differentiation of human ES cells into human A-NPC and P-NPC cells. We followed the protocol previously described[63]. Human ESC line H9 (WA09 line, NIH registry 0046, University of Wisconsin) was plated in clumps in the presence of Y27632 (10 mg ml^{-1}, 1,000×, Axon MedChem). The hES colonies were allowed to grow in mTESR (Stem Cell Technologies, 85850) for another day before the media was switched to differentiation media (1× N2, 1× B27, DMEM/F12:neurobasal = 1:1 (Invitrogen), 0.1 mM ascorbic acid (Sigma-Aldrich)) with small molecules (SB431542, LDN193189 and CHIR99021 (final concentration: 10 µM, 100 nM and 3 µM) from Stemgent and Tocris). To obtain anterior or posterior neural stem cells, SB431542/LDN194189 and SB431542/LDN193189/CHIR99021 were added, respectively. The cells were allowed to differentiate in the media for 6 d before they were dissociated with Accutase and plated at the density of 1 million cells per well of six wells for infection. For Ngn2 day 2 postdox induction cells, TetO-Ngn2-t2a-puro and FUW-rtTA were infected at day 0 in differentiation media followed by adding doxycycline and puromycin the next day. The cells were selected for 2 d in puromycin before collection for downstream experiments.

Generation of iN cells for psych MPRA. The inducible Ngn2 human ES cell line (H9) was created by infecting human ESC line H9 (WA09 line, NIH registry 0046) with TetO-Ngn2-PGK-puromycin^R and FUW-rtTA-IRES-hygromycin^R and selecting with puromycin and hygromycin for more than 1 week. The Ngn2 inducible line was maintained under feeder-free conditions in mTeSR (Stem Cell Technologies, 85850). For neuron generation, the inducible Ngn2 ES cell line was treated with Accutase and incubated for ~5 to 10 min at 37 °C to obtain single-cell suspension; resuspended 2×10^6 cells in 2 ml mTseR + Thiazovivin, and infected with 15 µl Psych MPRA

lentiviruses and plated on the Matrigel (BD Biosciences). After approximately 16–18 h of infection, the media was changed to N3 medium (DMEM/F12 1:1, 11320-033 (500 ml), 5 ml N2 (1×) 5 ml 17502-048, NEAA (1×), 1.6 ml insulin (stock 6.25 mg ml^{-1}, Sigma-Aldrich) 2.5 ml penicillin–streptomycin) with Dox to induce the TetO genes expression. On day 1, a 48 h puromycin selection (2 mg l^{-1}, Sigma-Aldrich), hygromycin (400 µg ml^{-1}) and blasticidin (2 µg ml^{-1}) period was started, and the medium containing antibiotic and Dox was changed every day.

MPRA library sequencing and analysis. Both the plasmid library and cDNA libraries for replicates of each cell condition were sequenced. The plasmid library was sequenced to assess oligo frequencies. Briefly, final library concentrations were assessed by Kapa Library Quantification Kit (Roche) and sequenced on an Illumina MiSeq instrument using 30 cycles for read 1 (barcode) and 115 cycles for read 2 (genomic instance). Paired-end sequencing was used to assess template switching. To count the number of reads per unique barcode sequence, we took the read 1 sequence, extracted the 20 bp region corresponding to the random barcode, and aligned the sequence to the reference MPRA library using bowtie2 (ref. 64). Barcodes were allowed a maximum of one mismatch during alignment. The barcodes were counted to determine barcode-oligo frequencies in the plasmid library. Similar alignment was done with read 2 to align the sequence to the genomic instance of the variant to assess for template switching.

cDNA libraries were synthesized from total RNA, as previously described[65]. Library concentrations were determined using a Kapa Library Quantification Kit (Roche) and on an Agilent Bioanalyzer. Average molarity was used to equally mix a final cDNA library. Libraries were sequenced on an Illumina MiSeq instrument to ensure each sample had adequate coverage. Deeper sequencing runs were performed on an Illumina NovaSeq 6000 SP flow cell (Novogene) with 50 read 1 cycles, 8 index 1 cycles and 20 index 2 cycles. cDNA library barcode-oligo counts were determined in the same manner as the DNA plasmid library.

ChIP PCR. For episomal ChIPs, ~18×10^6 HEK293T cells were transfected with 15 µg plasmid encoding 461 bp chromosomal fragments with the reference or alternative alleles of rs301806 and rs301807 in central position. Cells were collected ~20 h post-transfection and crosslinked with 1% formaldehyde. After nuclei isolation, samples were sonicated for 35 min using 30 s on/off cycles in Bioruptor (Diagenode) to fragment the plasmids. Samples were then immunoprecipitated with Normal Rabbit IgG (Cell Signaling Technology, 2729sor rabbit anti-RUNX1/AML1 antibody (Abcam, ab23980). Five microliter RUNX1 antibody (1 µg $µl^{-1}$) and 5 µl normal rabbit IgG antibody (1 µg $µl^{-1}$) were used for each ChIP reaction. Primer sets used for qPCR are listed in Supplementary Table 15.

CRISPRi. dCas9-BFP-KRAB from pC13N-dCas9-BFP-KRAB, a gift from Martin Kampmann (Addgene plasmid, 127968) was cloned into pLex-CMV. Guides were designed using CRISPick[66] and CRISPOR[67] to choose two guides for each target. Safe targeting guides were chosen from random region targeting sequences in ref. 68. CRISPRi guides were cloned into the pLentiGuide plasmid (Addgene, 117986) using Gibson assembly (NEB). The virus was made for both pLex-dCas9-BFP-KRAB and the guide library by transfecting LentiX 293T (Takara Bio) cells with Lipofectamine 3000 (Invitrogen). Neuroblastoma cell line SH-SY5Y and 293T cells were infected with concentrated pLex-dCas9-BFP-KRAB virus and guide virus. After infection, cells were selected with blasticidin and puromycin for 2–3 d. Cells were then plated into six-well plates and six replicates were collected for each guide. RNA was extracted using a Qiagen RNeasy kit. cDNA was prepared using iScript cDNA synthesis kit (BioRad) with 500 ng RNA. qPCR was performed for genes nearby or implicated by GTEx, eQTLgen or HiChIP using SYBR green PCR Master Mix (Thermo Fisher Scientific).

Homologous directed recombination (HDR) vector construction. The donor ssAAV vector used as a template for HDR was constructed by cloning of ~1,200 bp genomic DNA sequence flanking the 5′ and 3′ ends of the target SNP rs12264415 into the AAV transfer plasmid between AAV ITR sequences. For each editing experiment, two consecutive constructs were generated differed by the presence of the reference and alternate allele. During genomic amplification, T/G point mutation was engineered representing substitution of the reference allele T to the alternate allele G. For genomic DNA amplification, CNNM-F/R primers (Supplementary Table 15) contained homology arms to the AAV transfer vector allowing in-fusion assembly into NheI/EoRI digested plasmid. After confirmation of the insert sequence integrity for both reference and alternate allele-containing plasmids, constructs were used for AAV virus production at the Stanford Neuroscience Gene Vector and Virus Core. AAV-DJ serotyped donor ssAAV virus was produced at a genomic titer of 2 to 3×10^{13} transduction units (TU) ml^{-1} and used for HDR experiments at a multiplicity of infection (MOI) 2.5×10^5.

CRISPR and AAV-mediated HDR. The guide sequences targeting reference SNP for CRISPR/Cas9 or CRISPR/Cpf1 genome editing were predicted using CHOPCHOP web tool[69] and were ordered from IDT as sgRNA or crRNAs accordingly (see Supplementary Table 15 for gRNAs). For CRISPR/Cas9 genome editing, 73 pmol of the sgRNA was complexed with 61 pmol Recombinant Alt-RspCas9 protein (IDT), while for CRISPR/Cpf1 editing 150 pmol crRNA was complexed with equimolar amount of Alt-RAsCas12a (IDT) in 10 µl of Amaxa nucleofection buffer from kit V (Lonza) for 10 min and immediately used for nucleofection of 8×10^5 primary SH-SY5Y cells with Amaxa nucleofection apparatus (Lonza) using program G-004. After recovery, cells were mixed with AAV virus containing either reference or alternate allele containing donor template at MOI 2.5×10^5, split into two wells of a six-well plate and propagated for 72 h. After the culture reached 60–80% confluence, cells were transferred into 10-cm plates. Genomic DNA was then isolated and evaluated for editing efficiency using PCR amplification and sequencing of the bulk cell population. In addition, the amplified fragment was cloned into pBluescript vector, and the editing efficiency was evaluated through individual colony sequencing. The population of cells with at least over 70% HDR editing efficiency was used for further experiments.

Epigenomic data generation and processing. Detailed data generation, sequencing reads processing methods and differential analysis for RNA-seq, ATAC–seq and HiChIP are listed in Supplementary Methods. Briefly, RNA-seq data were generated using the Illumina protocol (015.96) for total RNA. RNA-seq reads were aligned to hg19 reference genome using STAR aligner (version 2.5.4b)[70] and TPM values were calculated using RSEM (version 1.3.0)[71]. Fast-ATAC sequencing on AST biological replicates was performed as previously described[72]. ATAC–seq read alignment, quality filtering, duplicate removal, transposase shifting, peak calling and signal generation were all performed through the ENCODE ATAC–seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline). HiChIP data generation protocol was performed for AST, ESC cells, N-D2, N-D4, N-D10 and N-D28 as previously described[18]. HiChIP paired-end reads were aligned to the hg19 genome using the HiC-Pro pipeline[73] v2.11.1, with additional filtering steps using hichipper[74] v0.7.0 and FitHiChIP[75].

ATAC–seq footprinting. ATAC–seq peaks were footprinted using the rgt package[76] (https://github.com/CostaLab/reg-gen), which uses an HMM-based model to identify active TF binding sites in open chromatin (aka ATAC peaks). HOCOMOCO v11 motifs used in this analysis were preprocessed to increase motif quality. Noninformative bases (information content < 0.4) were removed from both ends of the motif. To perform the footprinting process, bam files for each cell type were merged and sorted into a single file. Similar merging and sorting

process is performed for bed files for each cell type. 'rgt-hint footprinting' was called for each cell type. And then 'rgt-motif analysis matching' was called to identify HOCOMOCO motif PWMs (https://hocomoco11.autosome.org/) that map to the specified footprinted region.

Target gene identification. Target genes of daSNV-associated TFs were determined using chromatin interaction (HiChIP) and accessibility (ATAC–seq) data gathered within the H9-derived cell model. First, TFs in the HOCOMOCO v11 database (https://hocomoco11.autosome.org/) were identified via DNA footprinting. A target gene of a daSNV TF is defined as having an ATAC footprint of the TF's corresponding motif within the target gene's promoter region (defined as 200 bp upstream and 50 bp downstream) of the transcription start site. Distally looped target genes (where the TF footprint was brought into the proximity of the target gene's TSS), were analyzed but not included in the final analysis. Additionally, both the TF and target gene had to express TPM > 1 in cell type matching the ATAC profile. GO enrichment analysis was performed using the clusterProfiler[77] 3.14.0R package, and visualized using ggplot.

Enrichment calculation. To determine the GO processes enriched in daSNV-associated genes versus genes from our initial GWAS database, gene set enrichments were calculated using a hypergeometric test using the R package ClusterProfileR[77] v3.14.0, and significance was reported for FDR-corrected $P < 0.05$. Enrichments were mostly shown as dotplots with gene set (that is GO biological process, REACTOME pathway, etc.) on one axis and cell type on the other axis. The size of the circle represents the number of genes associated with a given gene set. The color of the circle represents the multiple hypothesis corrected P value for the hypergeometric test.

Cell-specific heritability estimates. Stratified LDSC, as previously described[25–27], can be used to obtain cell-specific disease heritability estimates using GWAS summary statistics. Preprocessed summary statistics from 25 neuropsychiatric GWAS and 19 non-neuropsychiatric GWAS were obtained from the Broad Institute webpage (https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/). Using the LDSC[28,78] software provided on GitHub (https://github.com/bulik/ldsc) and reference data on the Broad Institute webpage (https://alkesgroup.broadinstitute.org/LDSCORE/), annotation and LD score files were generated for each cell type from ATAC–seq and RNA-seq from this study with a control epigenetic set extracted from previously published epithelial dataset[79].bed files containing cell-specific epigenetic and transcriptomic profiles, separately. Using standard parameters, the 'make_annot.py' and 'ldsc.py' (with the '–l2' flag) scripts were first used to generate the cell-specific annotation and LD files, then the 'ldsc.py' (with the '--h2-cts' flag) script was used to generate cell-type-specific portioned heritability scores for each GWAS.

eQTL-based gene networks. Gene networks for each disease were generated by extracting eQTLs listed in GTEx v8 (for neural tissues) and PsychEncode eQTLs for daSNVs of each disease and then plotted in CytoScape v3.7.2. Genes were linked via STRING-db[80] and color-coded based on annotated protein function. Function clusters were curated based on frequently appearing key terms. Putative disease-specific and psychiatric genes of interest were annotated based on a PubMed API query for the (psychiatric disease terms) AND (gene of interest).

Brainmap SMART-seq cortical data[81] were used to annotate eQTL-based gene sets, and heatmaps depicting gene expression across the 128 cortical cell types are generated via R package pheatmap.

Mendelian-polygenic disease association analysis. To generate a database of neurogenic Mendelian diseases, we curated neuro-related conditions from the Online Mendelian Inheritance in Man (OMIM) database[59], yielding 68 neuro-relevant Mendelian (or rare) diseases and 1,132 genes (Supplementary Table 12). Diseases were chosen based on whether a CNS-related symptom was central to the disease phenotype as the primary mechanism of disease. A hypergeometric test was used to assess the significance of the genes that overlap between neuro-relevant Mendelian disease genes and daSNV gene MPRA hits. A Venn diagram was plotted using the Python venn2 function. Gene set enrichment analysis was performed on the overlapping genes using EnrichR[82] (https://maayanlab.cloud/Enrichr/). GWAS Catalog 2019 and Human Gene Atlas gene set enrichment results were plotted as bar plots.

Protein coding mutation analysis. Protein coding variants for SCZ were extracted from SCHEMA[83] (https://schema.broadinstitute.org/). Proteins were a SCHEMA meta-analysis $P < 0.05$ were shown. The associated proteins intersected with the SCZ eQTL gene list and SNVs were visualized in the WashU Epigenome Browser to determine whether the daSNVs link to the gene of interest. A list of seven proteins was curated based on the existence of protein-truncating variant present and daSNV–gene linkage present in chromatin looping data.

UK Biobank analysis. From the UK Biobank (http://www.nealelab.is/uk-biobank/), we extracted the GWAS results for 806 psych codes (Supplementary Table 14). We used LD score regression software (https://github.com/bulik/ldsc) to determine phenotype heritability (see Cell-specific heritability estimates for implementation details). Additionally, we colocalized the UK Biobank SNVs with the daSNV hits from MPRA and filtered for significant SNVs using an FDR cutoff of 0.10. Subsequently, we generated a heatmap using pheatmap, showing log-normalized β values for genes associated with the daSNV of interest and a psychiatric clinical phenotype. We collapsed the 806 psych codes by phenotype, based on a similarity of their β-normalized scores across the daSNVs and manually annotated phenotype clusters ($n = 64$). The resultant heatmap was further collapsed as only one gene/per SNV was used to represent the clinical phenotypic profile shown, based on prioritization schema favoring genes with brain-specific expression modulated by, chromatin linkage to, and prior literature evidence for the given daSNV.

Predicting chromatin effects on noncoding sequence variants. To observe how MPRA-derived allele-specific variants compare to existing sequence-based computation algorithms for predicting noncoding variant effects, the 2,221 MPRA variants were run against DeepSea[16] (available on http://deepsea.princeton.edu/job/analysis/create/) and gkmSVM[84] v0.82.0 as described in the published R package instructions. The distribution of score differences between the reference and alternate sequences for daSNVs versus non-SNVs were compared using a Student's t test for each cell type.

Reporting summary

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

Data availability

All raw and processed sequencing data are available in GEO accession GSE182095. For ease of reference, processed TPM values for RNA-seq are provided in the data supplement (Supplementary Table 13). Tracks for ATAC and HiChIP were visualized on WashU Epigenome Browser. Raw and processed RNA, ATAC and HiChIP D0 and D2 Ngn2-derived H9 samples are referenced[85]. All MPRA summary statistics and raw count results are provided (Supplementary Data 5) and processed data is available at https://arvid-data.shinyapps.io/neuropsychiatry/ or Supplementary Data 3. Previously published GWAS study data used as a basis for this work is noted in each study published and is annotated in online GWAS data resources, including https://www.ebi.ac.uk/gwas/ and http://www.nealelab.is/uk-biobank/. For LDSC scoring, GWAS data used was available and preprocessed by https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/. For colocalization

studies, available summary statistics are provided: https://zenodo.org/record/3518299#.XbMgFNF7m90. Additional publicly available data sets used include GTEx v7, Haploreg v4, ENCODE hg19, StringDB (https://string-db.org/), OMIM (https://www.omim.org/), The Drug Repurposing Hub (http://www.broadinstitute.org/repurposing), PsychENCODE (http://resource.psychencode.org/), HOCOMOCO v11 (https://hocomoco11.autosome.org/), UCSC browser (https://genome.ucsc.edu/), Brainmap SMART-seq cortical data (http://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq), SCHEMA (https://schema.broadinstitute.org/) and SNVlocs.Hsapiens.dbSNV142.GRCh37. VA cohort data have restricted access due to privacy concerns.

Code availability

Analyses were done in custom Jupyter Notebook or Rmarkdown scripts in Python 3.7.4 and R 3.6.1, locally or on the Stanford Sherlock computing cluster. Code to analyze transcriptomics and epigenomics data is available on GitHub (https://github.com/mguo123/pan_omics_psych.git)[86]. MPRA-based analysis scripts are available here (https://github.com/mguo123/psych_mpra.git)[87]. Additional software used includes LDSC (LD Score) (v1.0.1), MPRAnalyze (v1.4.0), STAR aligner (v2.5.4b), RSEM (v1.3.0), ENCODE ATAC−seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline), Bowtie2 (2.3.4.1) and EnrichR (https://maayanlab.cloud/Enrichr/). ChIPSeeker (v1.22.0), motifBreakR (v2.10.2), rgt (https://github.com/CostaLab/reg-gen), ClusterProfileR (v3.14.0), RColorBrewer (v1.1.0) HiC-Pro (v2.11.1), Hichipper (v 0.7.7), FitHiChIP (v7.0.0), diffloop(v1.10.0), DESeq2 (v1.26.0), CytoScape v3.7.2, ABC-Enhancer-Gene-Prediction (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction), gatk (v4.1.9.0), picard (v2.24.0), MACS2 (v2.1.1), enloc (https://github.com/xqwen/integrative), PhenomeXcan (https://github.com/hakyimlab/phenomexcan), gkmSVM (v0.82.0), DeepSea (http://deepsea.princeton.edu/job/analysis/create/), pheatmap (v1.0.12), biothings (v0.2.6) and GenomicRanges (v.1.48.0), and Rsubread (v2.0.0).

References

57. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
58. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
59. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
60. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–D881 (2016).
61. De Souza, N. The ENCODE project. *Nat. Methods* **9**, 1046 (2012).
62. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
63. Du, Z. W. et al. Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells. *Nat. Commun.* **6**, 6626 (2015).
64. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
65. Kim, D. et al. The dynamic, combinatorial *cis*-regulatory lexicon of epidermal differentiation. *Nat. Genet.* **53**, 1564–1576 (2021).
66. Kim, H. K. et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).
67. Concordet, J. & Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, 242–245 (2018).
68. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
69. Labun, K., Montague, T. G., Krause, M., Cleuren, Y. N. T. & Valen, E. CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, 171–174 (2019).
70. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
71. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
72. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
73. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
74. Lareau, C. A. & Aryee, M. J. Hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat. Methods* **15**, 155–156 (2018).
75. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. FitHiChIP: identification of significant chromatin contacts from HiChIP data. Preprint at *bioRxiv* https://doi.org/10.1101/412833 (2018).
76. Li, Z. et al. Identification of transcription factor binding sites using ATAC–seq. *Genome Biol.* **20**, 45 (2019).
77. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
78. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
79. Donohue, L. K. H. et al. A *cis*-regulatory lexicon of DNA motif combinations mediating cell-type-specific gene regulation. *Cell Genom.* **2**, 100191 (2022).
80. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
81. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
82. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
83. Singh, T. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* **604**, 509–516 (2022).
84. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol.* **10**, 1003711 (2014).
85. Ang, C. E. et al. The dynamic interplay between homeodomain transcription factors and chromatin environment regulates proneural factor outcomes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.02.398677 (2020).
86. mguo123. mguo123/pan_omics_psych: publication release. Zenodo. https://zenodo.org/record/8098407 (2023).
87. mguo123. mugo123/psych_mpra: publication release. Zenodo. https://zenodo.org/record/8098409 (2023).

Acknowledgements

We thank S. Srinivasan for the creation of the web resource. We thank G. Rayant and K. Fields for their generous support and helpful discussions. We also thank M.P. Snyder, W.J. Greenleaf, D.F. Levinson and H.Y. Chang for the presubmission review and J. Engreitz and members of the Khavari and Altman laboratories for helpful discussions. This work was supported by the USVA Office of Research and Development and by the Atlas of Regulatory Variants in Disease (ARVID) project from NHGRI/NIH U24HG010856 and by NIAMS/NIH AR076965 (both to P.A.K.). S.B.M. was supported by R01AG066490 and R01MH125244. A.E.U. was supported by P50HG00773506. This work was supported by Bitscopic's R&D budget and intramural funding

from the Department of Veterans Affairs. The funders had no role in study design, data collection and analysis, the decision to publish or the preparation of the manuscript. We thank C.C. Lee, H. Parekh and J. Mewton for Praedico maintenance, assistance with data extraction and professional support.

Author contributions

M.G.G., M.W. and P.A.K. conceptualized the project. M.G.G., D.L.R., C.E.A., Y.L., Y.Y., Y.Z., L.N.K., L.K.H.D., X.Y., L.M., T.F., I.E., A.H., Z.S., Y.P., V.B. and N.A. performed experiments and analyzed the data. M.G.G., P.A.K., M.W., R.B.A., L.E.D., A.E.U., S.B.M., P.E., M.H. and D.H.G. guided methodology development, experiments and data analysis. M.G.G. and P.A.K. wrote the manuscript with input from all authors.

Competing interests

V.B. and P.E. are employees of Bitscopic. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01533-5.

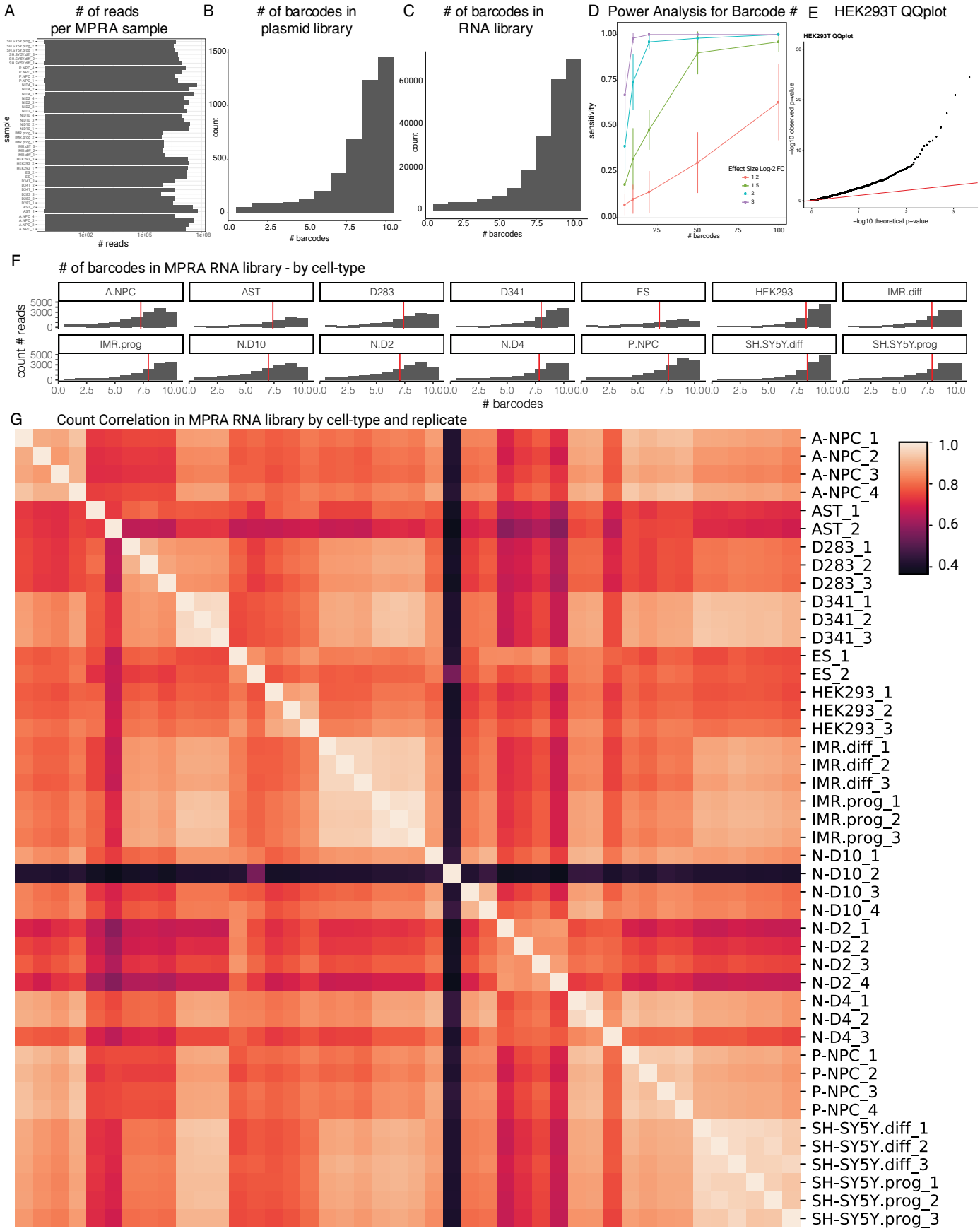
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01533-5.

Correspondence and requests for materials should be addressed to Paul A. Khavari.

Peer review information *Nature Genetics* thanks Yin Shen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

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

https://doi.org/10.1038/s41588-023-01533-5

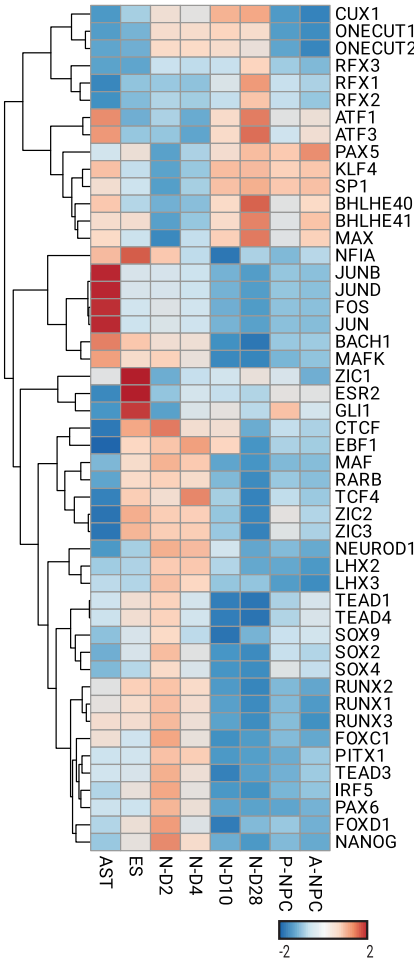


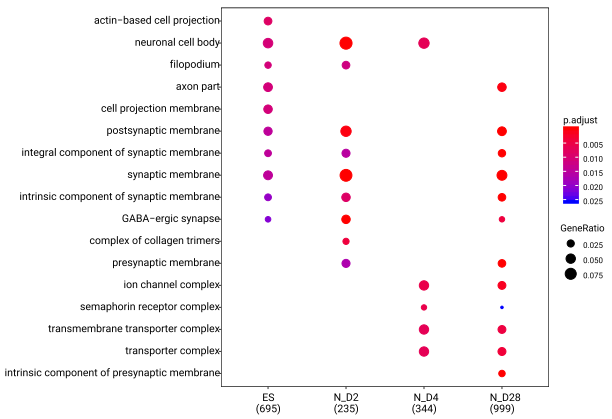
Extended Data Fig. 1 | See next page for caption.

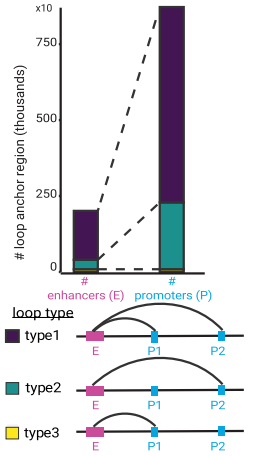
https://doi.org/10.1038/s41588-023-01533-5

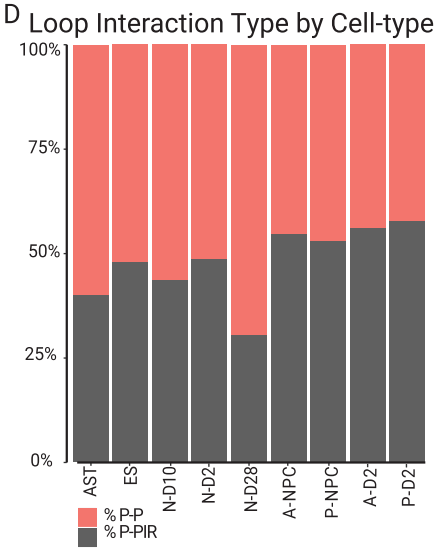
Extended Data Fig. 1 | MPRA QC statistics. (a) Bar chart showing number of reads per MPRA sample in log scale. Replicates are the number following the "R" prefix. Cell type abbreviations are as follows: AST= astrocytes; ES=hESC or human embryonic stem cell; A-NPC = anterior neural progenitor cell; P-NSC = posterior neural progenitor cell; N-DX = induced neuron of day X. Histograms showing barcodes per sequence in the **(b)** plasmid (prior to lentiviral infection) and **(c)** RNA library (extract post infection). **(d)** Power analysis for different levels of barcodes power for at 4 different log2-fold change thresholds (1.2, 1.5, 2, 3), for a total of 20 variants, simulated 5 times with 5, 10, 20, 50, and 100 barcodes each. **(e)** QQ plots showing the -log10 empirical vs theoretical p-values derived from MPRAnalyze for HEK293T as a cell-type example. The red line is (y = x). **(f)** Histograms showing barcodes per sequence in the RNA library, by cell type. **(g)** Heatmap showing Pearson count correlation between replicates for all cell types, conditions, and replicates.

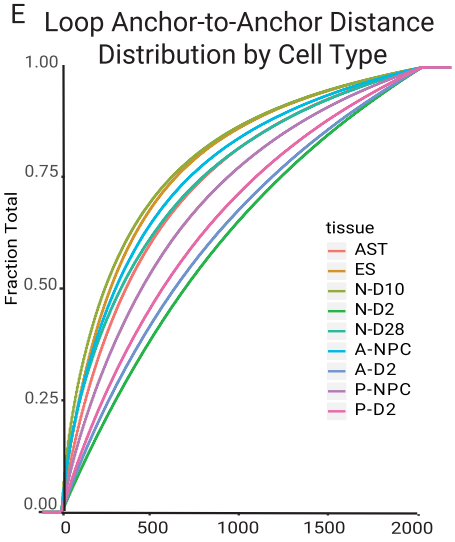
https://doi.org/10.1038/s41588-023-01533-5

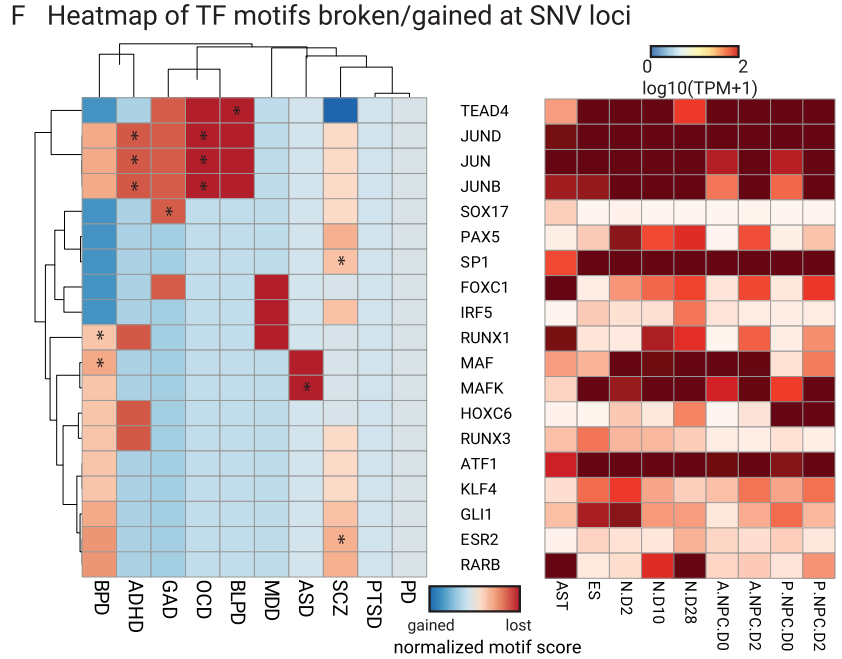


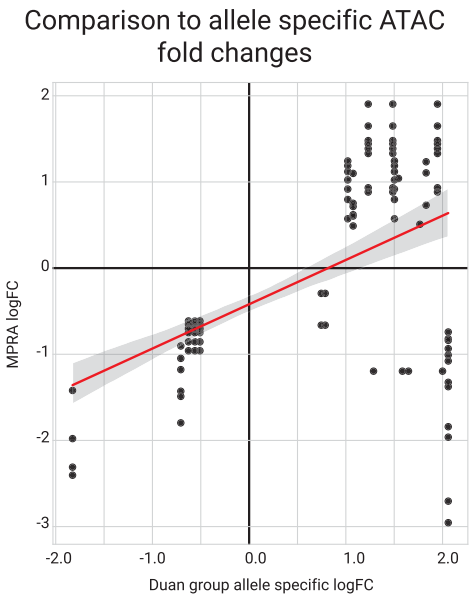










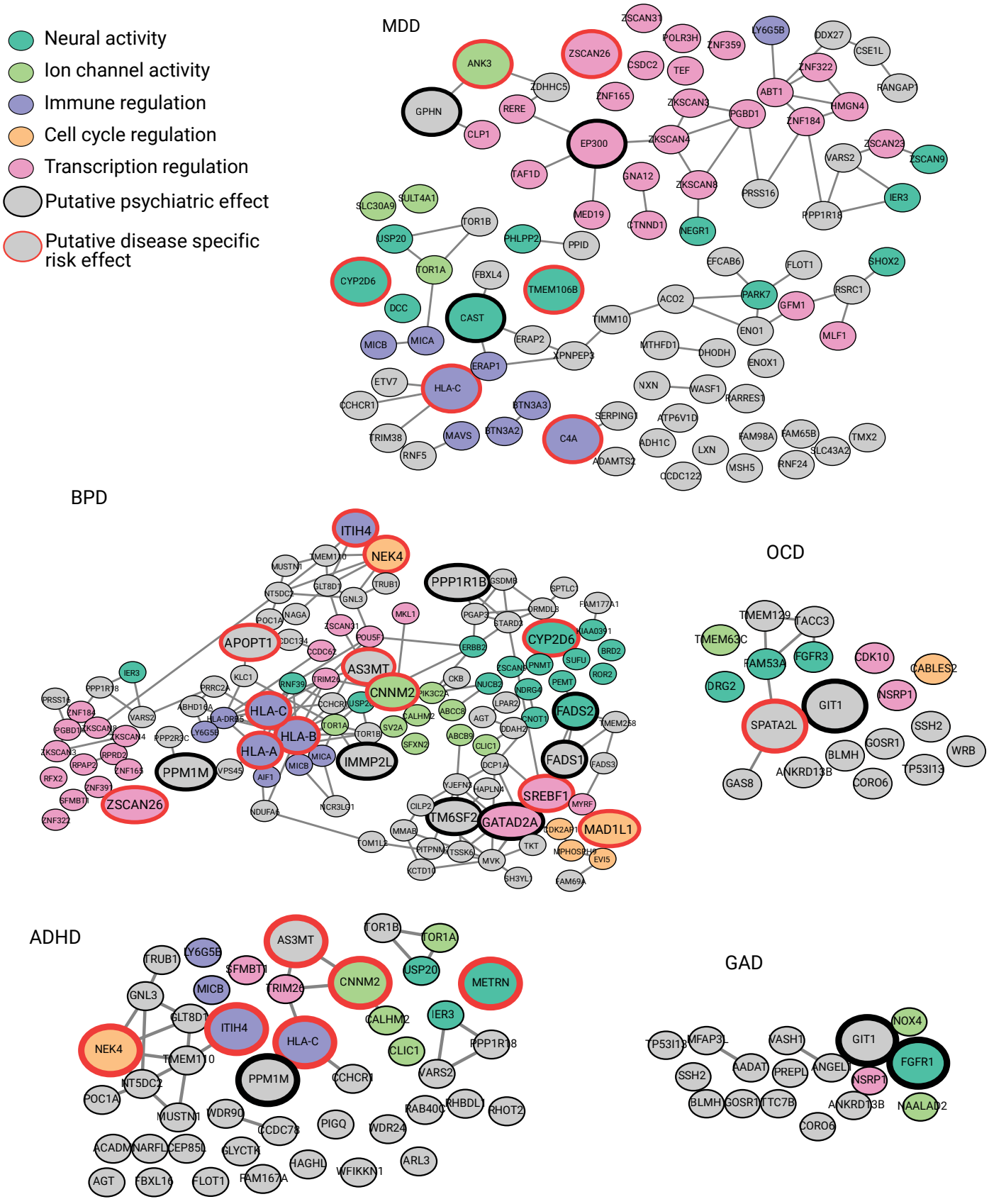


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41588-023-01533-5

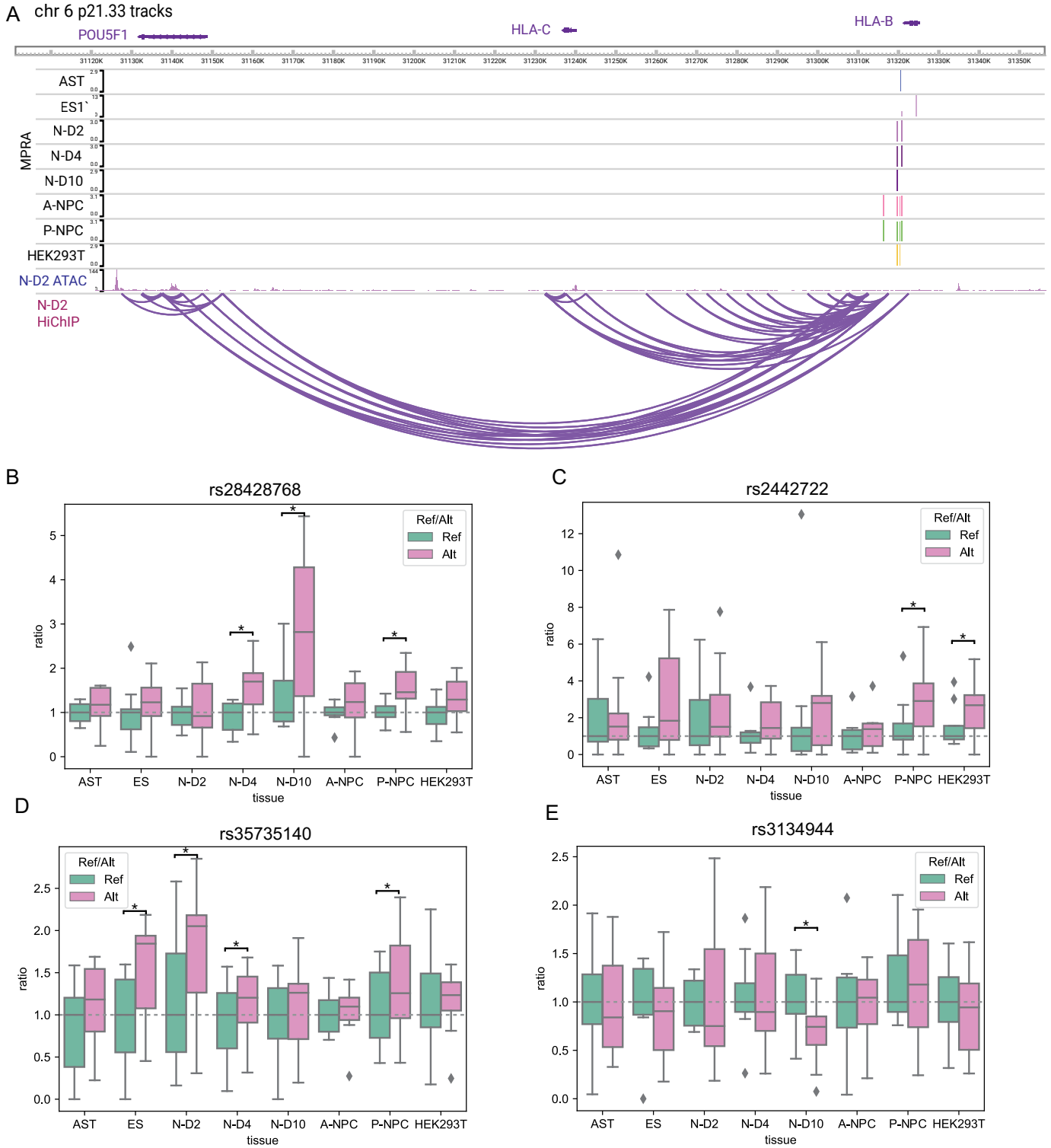
Extended Data Fig. 2 | Epigenetics study of the role of transcription regulation in neuropsychiatric diseases. (a) Heatmap showing TF footprints that are enriched in cell types; color scale is normalized count values. **(b)** GO biological process dotplot depicting enrichment terms for genes closest to ATAC accessible peaks found across ES-derived neuronal differentiation. The size of the dot is the number of genes in the GO geneset and the color indicates FDR-adjusted p-values. **(c)** Bar chart showing frequency of loop types in promoters and promoter interaction anchor loops (putative enhancers) derived from HiChIP data. Type 1: where an enhancer is linked to a distal gene and the nearest gene, Type 2: where an enhancer is linked only to a distal gene, Type 3: where an enhancer is looped to the closest gene. **(d)** % of P-P (promoter-promoter) and P-PIR (promoter to promoter interaction regions) loops per cell type found via HiChIP. **(e)** Cumulative distribution curves of distance between loop anchors for the different tissues. **(f)** Heatmap (left) showing normalized enrichment scores of motifs broken or gained by SNVs associated with different neuropsychiatric diseases derived from MotifBreakR, relative to a background of other neuropsychiatric diseases. The * refers to motifs that are significantly broken (p-value < 0.10, Fisher's exact test) in daSNVs compared to non-daSNVs for a specific disease. The heatmap (right) shows the log TPM expression values of these transcription factors in different neuronal cell lines and cell lines. **(g)** Scatterplot comparing log-2 fold changes (n = 206 variants) for the MPRA dataset (y-axis) with an external[20] allele specific open chromatin dataset (x-axis), with a Pearson correlation of 0.48, p-value 1.7×10^{-13}.

https://doi.org/10.1038/s41588-023-01533-5



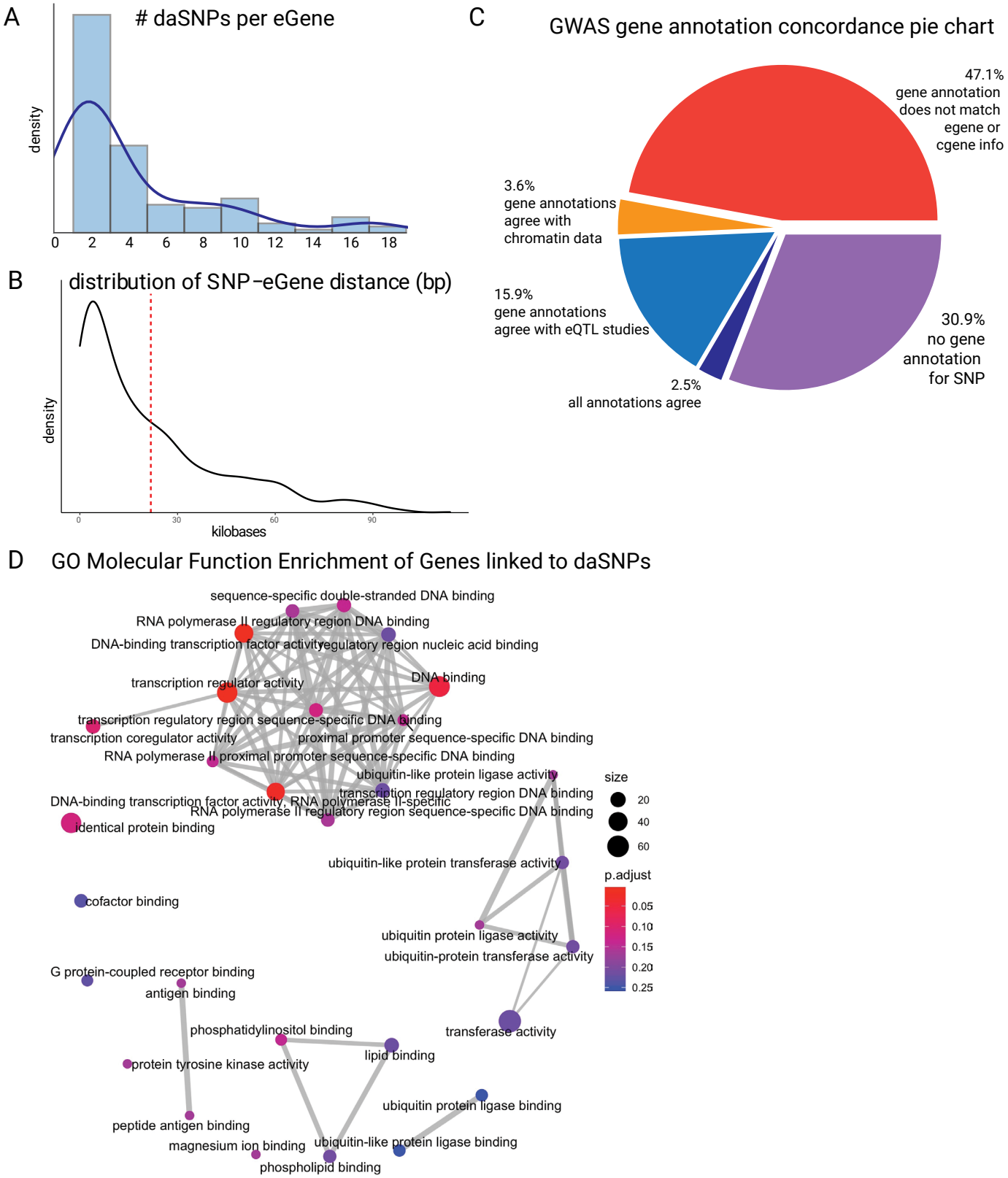
Extended Data Fig. 3 | eGene network analysis of additional diseases. eGene networks for the additional neuropsychiatric diseases with at least 20 eGenes (from left to right, top to bottom): MDD, BPD, OCD, ADHD, and GAD.

https://doi.org/10.1038/s41588-023-01533-5



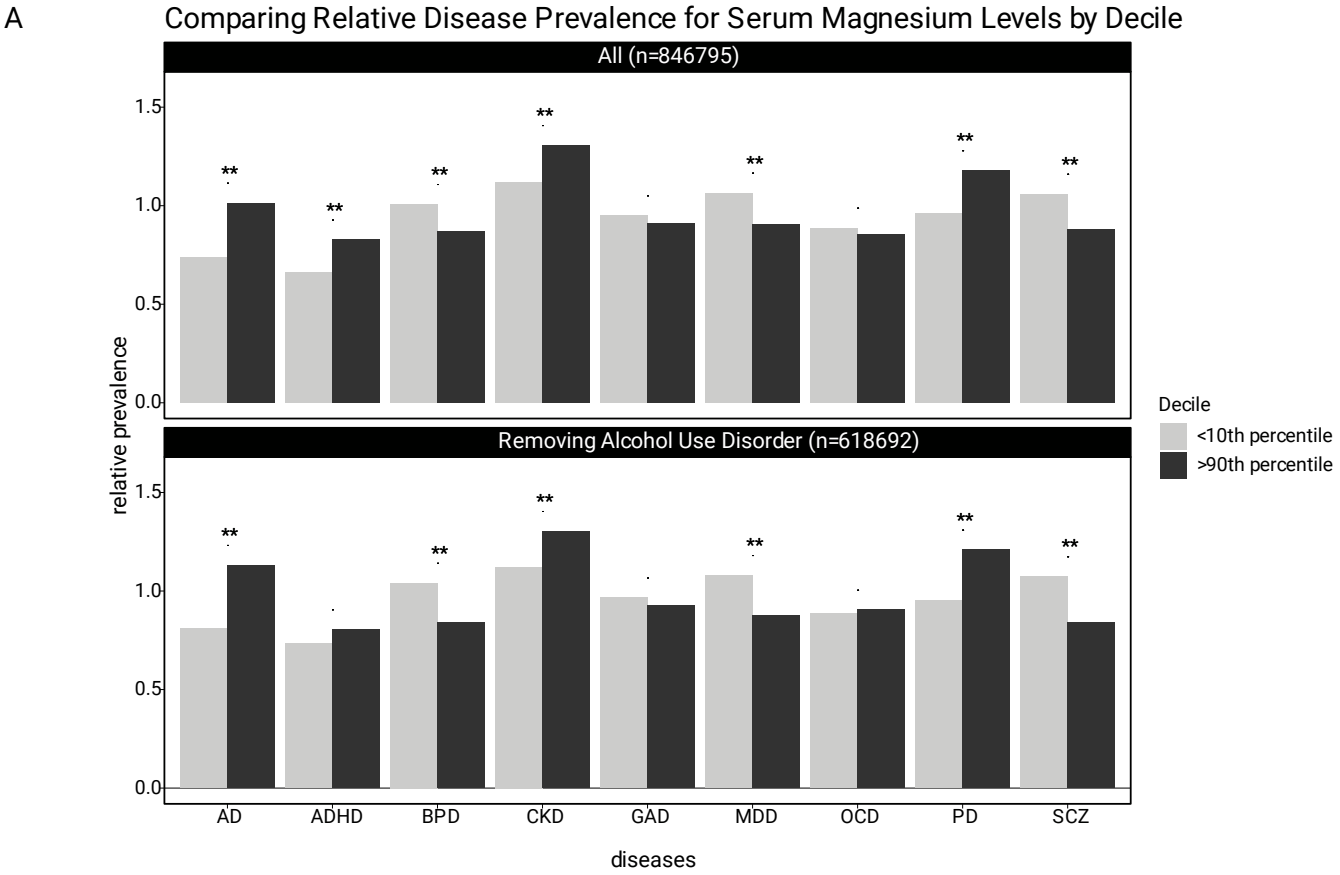
Extended Data Fig. 4 | *POU5F1/OCT4* vignette. (a) Tracks for the *POU5F1/OCT4* TF gene, where the peak tracks show the logFC change from cell-type specific MPRA for the daSNVs, and the bottom loop track shows the looping data for N-D2 cell type. Boxplots depicting ratios of cDNA to plasmid counts for reference versus alternate allele for SNVS **(b)** rs28428768, **(c)** rs2442722, **(d)** rs35735140, and **(e)** rs3134944, where the center line is the median of each MPRA normalized ratio (n = 10 genomic instances each); box limits are the upper and lower quartiles, whiskers are the 1.5x interquartile range, and points shown are outliers. Ratios are normalized to the median reference value for each cell type. Significant associations found by MPRAnalyze (FDR < 0.05) are shown with an asterisk*.

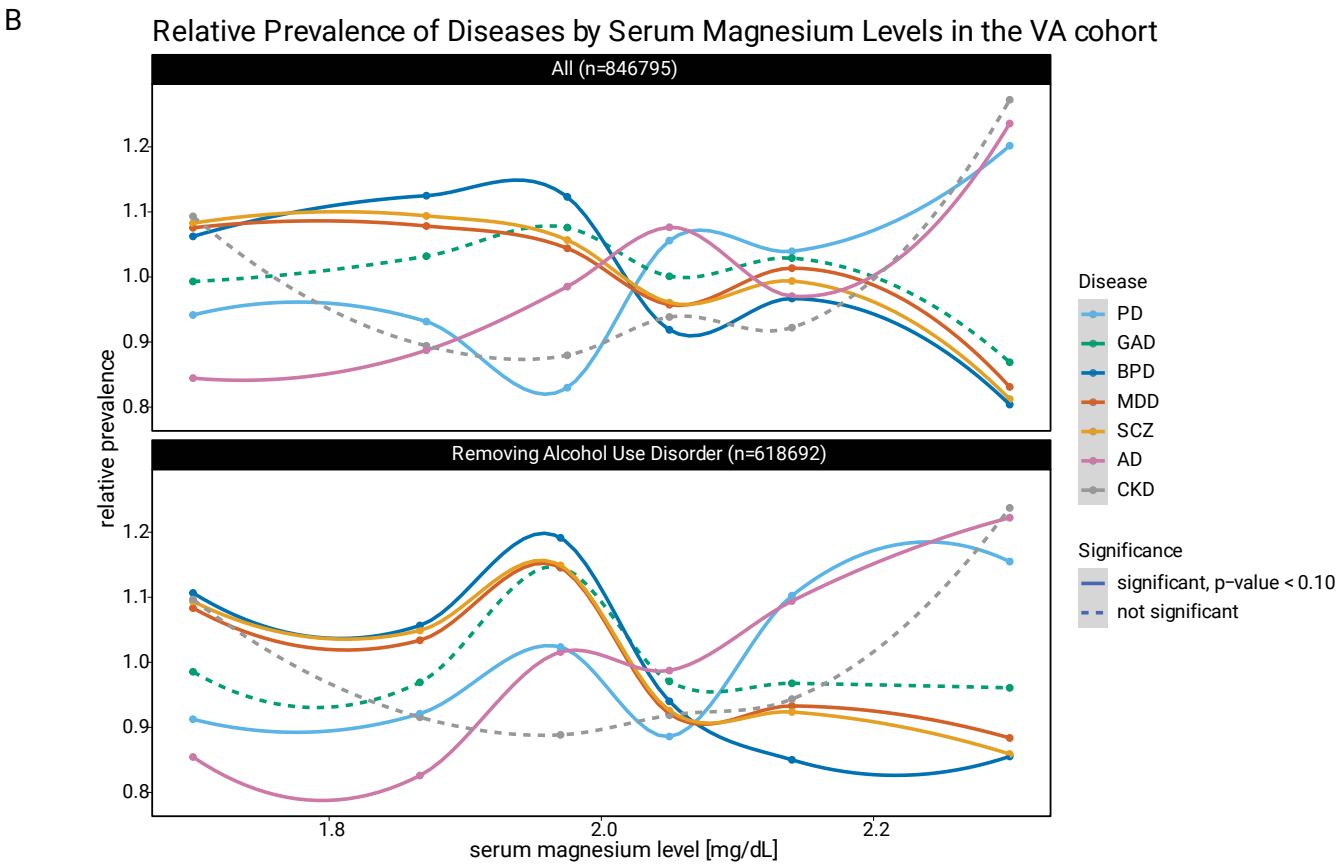
https://doi.org/10.1038/s41588-023-01533-5



Extended Data Fig. 5 | Gene concordance for variant annotation approaches. **(a)** Distribution of # daSNVs for a GTEx eGene annotations show eGenes are on average, linked to five daSNVs. **(b)** Density plot showing the distribution of daSNV-to-eGene distance with the mean depicted as a vertical red dotted line at 20kB. **(c)** Pie chart showing gene annotation concordance between the different annotation of daSNVs, indicating almost a half of GWAS gene annotations do not match expression or chromatin-based gene linkages. **(d)** Enrichment map made via ClusterProfiler showing GO Molecular Functions enriched in genes linked to daSNVs.

https://doi.org/10.1038/s41588-023-01533-5



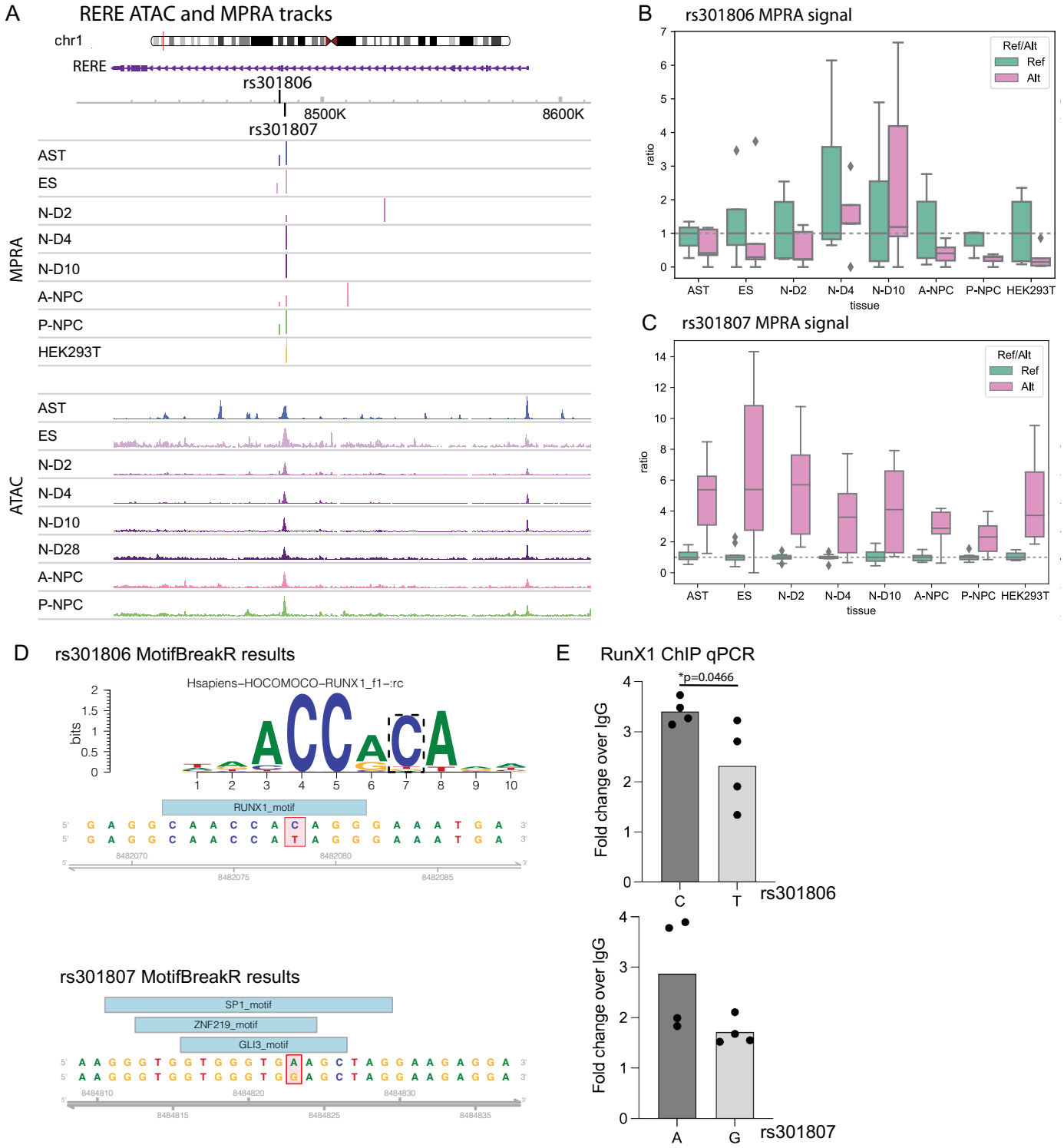


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41588-023-01533-5

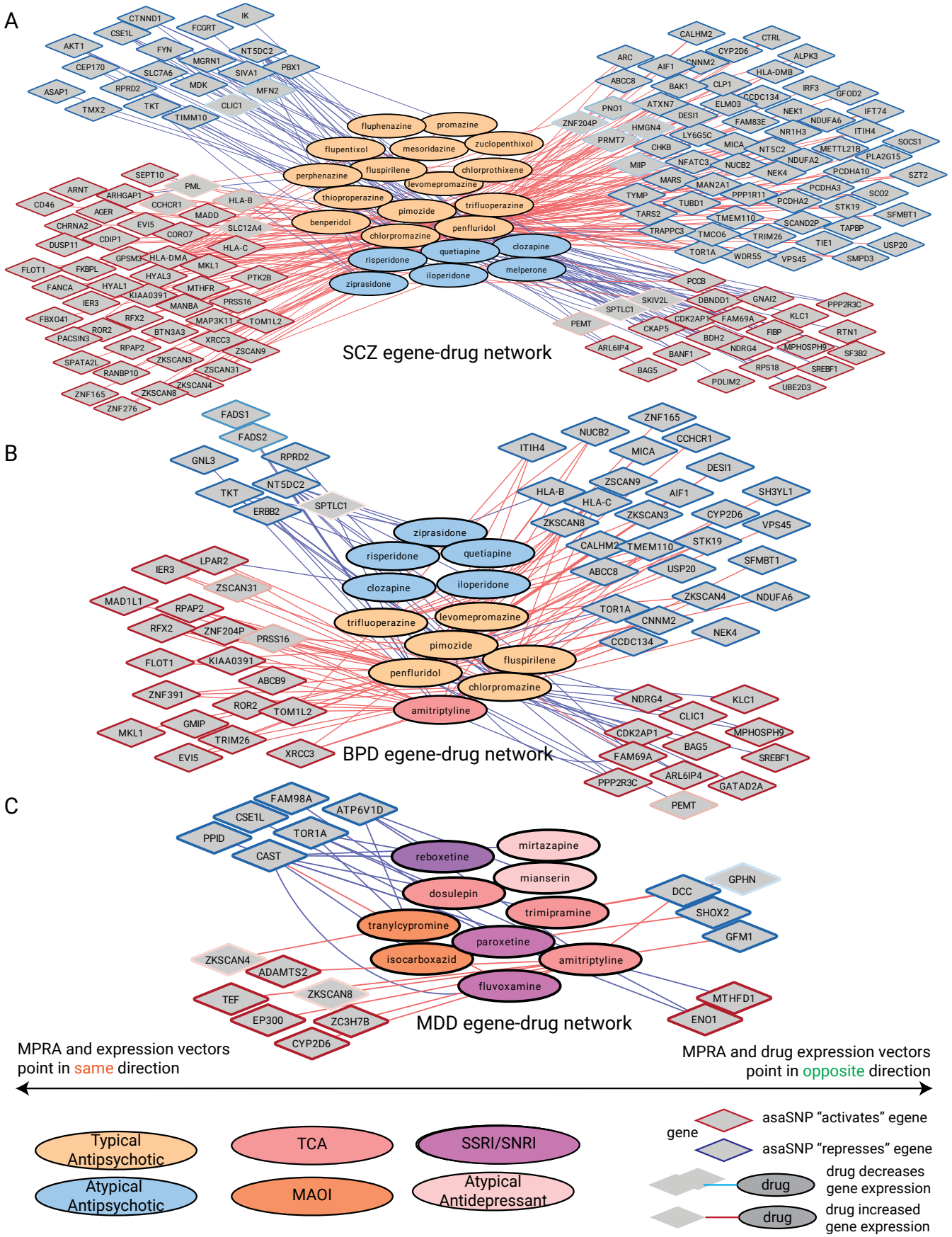
Extended Data Fig. 6 | Association between serum magnesium levels and relative psychiatric disease incidence in a VA cohort. (a) Relative disease prevalence for serum magnesium levels in the bottom 10th and upper 10th deciles. The 10th decile of serum magnesium are values < 1.6 mg/dL and the 90th decile of serum magnesium are values > 2.4 mg/dL. ** indicates significance between the two proportion based on a two-sided 2-proportion z-test FDR-corrected $p < 0.05$ for a given disease. **(b)** Relative prevalence of diseases by serum magnesium levels in the VA cohort. The above graph includes all patients age 45-85, n = 846795. The below graph removes all patients who were diagnosed with Alcohol Use Disorder, n = 618692. Cohort was partitioned by serum magnesium levels into 6 quantiles and the prevalence of each disease was calculated within the quantile. Relative prevalence is calculated as the prevalence normalized to the disease prevalence in the entire cohort. Significance is determined by linear regression with the null hypothesis beta = 0, with p-values < 0.10 shown in solid. Abbreviations of disease are as follows: ADHD = attention deficit hyperactivity disorder, PD = panic disorder, GAD = generalized anxiety disorder, BPD = bipolar disorder, MDD = major depressive disorder, OCD = obsessive compulsive disorder, SCZ = schizophrenia, AD = Alzheimer's disease, CKD = chronic kidney disease.

https://doi.org/10.1038/s41588-023-01533-5



Extended Data Fig. 7 | RERE vignette. (a) Tracks for gene *RERE*, where the MPRA peak tracks show the logFC change from cell-type specific MPRA for the daSNVs, and the bottom ATAC peak tracks show accessibility profiles for all cell types. Box-and-whiskers plots depicting ratios of cDNA to plasmid counts for reference versus alternate allele for daSNVs **(b)** rs301806, the SNV of interest and **(c)** rs301807, as comparison, where the center line is the median of each MPRA normalized ratio (each point is a genomic instance with at least one count), box limits are the upper and lower quartiles, whiskers are the 1.5x interquartile range, and points shown are outliers. Ratios are normalized to the median reference value for each cell type. Additionally, MotifBreakR results are shown for **(d)** rs301806 (above) and rs301807 (below), depicting loss of RUNX1 motif in rs301806, and no RUNX1 motif present at rs301807 loci. **(e)** ChIP PCR for the transcription factor RUNX1 with $n = 4$ replicates, * indicated significance of two-sided paired t-test p-value between the reference and alternate allele for the two SNPs.

https://doi.org/10.1038/s41588-023-01533-5



Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41588-023-01533-5

Extended Data Fig. 8 | CMAP drug perturbation analysis. Drug-eGene networks for **(a)** SCZ, **(b)** BPD, and **(c)** MDD. Linkages between eGene to drug indicate that the drug significantly upregulated (red) or downregulates (blue) the expression of that gene in neuro-relevant cell lines in CMAP. Genes (diamonds) are outlined based on the MPRA log fold change direction (red: positive, blue: negative). Drugs (ellipses) are color coded by drug type. Drug-gene pairs towards the left side of the map indicate the MPRA and expression vectors point in the same direction (putatively side effect causing variants); drug-gene pairs towards the right side of the map indicate MPRA and expression vectors pointing in the opposite direction (putatively therapeutic effects).

nature portfolio

Corresponding author(s): Khavari

Last updated by author(s): 8/7/2023

Reporting Summary

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

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection was either performed manually or using custom Jupyter notebook in Python 3.7.4, locally or on the Stanford Sherlock computing cluster. Data collection for transcriptomics and epigenomics data is available on GitHub (https://github.com/mguo123/pan_omics_psych.git, with publication release https://zenodo.org/badge/latestdoi/381854970). MPRA and SNP related data collection are available here (https://github.com/mguo123/psych_mpra.git, with publication release: https://zenodo.org/record/8098409).
Data analysis	Data analyses were done in custom Jupyter notebook or Rmarkdown scripts in Python 3.7.4 and R 3.6.1, locally or on the Stanford Sherlock computing cluster. Code to analyze transcriptomics and epigenomics data is available on GitHub (https://github.com/mguo123/pan_omics_psych.git, with publication release https://zenodo.org/badge/latestdoi/381854970). MPRA-based analysis scripts are available here (https://github.com/mguo123/psych_mpra.git, with publication release: https://zenodo.org/record/8098409). Additionally software used includes: LDSC (LD Score) (v1.0.1), MPRAnalyze (v1.4.0), STAR aligner (v2.5.4b), RSEM (v1.3.0), limma (v3.42.0), ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline), Bowtie2 (2.3.4.1), EnrichR (https://maayanlab.cloud/Enrichr/). ChIPSeeker (v1.22.0), motifBreakR (v2.10.2), rgt (https://github.com/CostaLab/reg-gen), ClusterProfileR (v3.14.0), RColorBrewer (v1.1.0) HiC-Pro (v2.11.1), Hichipper (v 0.7.7), FitHiChIP (v7.0.0), diffloop(v1.10.0), DESeq2 (v1.26.0), CytoScape v3.7.2 , ABC-Enhancer-Gene-Prediction (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction), gatk (v4.1.9.0), picard (v2.24.0), MACS2 (v2.1.1), enloc (https://github.com/xqwen/integrative), PhenomeXcan (https://github.com/hakyimlab/phenomexcan), gkmSVM (v0.82.0), DeepSea (http://deepsea.princeton.edu/job/analysis/create/), pheatmap (v1.0.12), biothings (v0.2.6), GenomicRanges (v.1.48.0), Rsubread (v2.0.0).

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All raw sequencing data are available in GEO accession #GSE182095. Processed TPM values for RNA seq is provided in the data supplement (Data S12). Tracks for ATAC and HiChIP can be accessed in WashU Epigenome Browser using the data supplement (Data S13-14). Additional processed peak, loop, transcript files can be found on the AWS server (s3://mguo123-washu/psych_mpra_paper_data/pan_omics_psych_data/data/raw), for ease of importing into the corresponding analysis code. Raw and processed RNA, ATAC, and HiChIP D0 and D2 Ngn2-derived H9 samples are referenced. All MPRA summary statistics and raw counts results are provided (Data S15). Previously published GWAS study data used as a basis for this work is noted in each study published and is annotated in online GWAS data resources, including https://www.ebi.ac.uk/gwas/ and http://www.nealelab.is/uk-biobank/. For LDSC scoring, GWAS data used was available and preprocessed by https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/. For colocalization studies, available summary statistics are provided: https://zenodo.org/record/3518299#.XbMgFNF7m90. Additional publicly available data sets used include: GTEx v7, Haploreg v4, ENCODE hg19, StringDB (https://string-db.org/), OMIM (https://www.omim.org/), The Drug Repurposing Hub (http://www.broadinstitute.org/repurposing), PsychENCODE (http://resource.psychencode.org/), HOCOMOCO v11 (https://hocomoco11.autosome.org/), GO (http://geneontology.org/), Reactome (https://reactome.org/), MSigDB (https://www.gsea-msigdb.org/gsea/msigdb), UCSC browser (https://genome.ucsc.edu/), Brainmap SMART-seq cortical data (http://portal.brain-map.org/atlases-and-data/rnaseq/human-multiple-cortical-areas-smart-seq), SCHEMA (https://schema.broadinstitute.org/), and SNVlocs.Hsapiens.dbSNV142.GRCh37.
VA cohort data has restricted access for privacy concerns.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Previously published GWAS study data used as a basis for this work is noted in each study published and is annotated in online GWAS data resources, including https://www.ebi.ac.uk/gwas/. For LDSC scoring, GWAS data used was available and preprocessed by https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/. For colocalization studies, available summary statistics are provided: https://zenodo.org/record/3518299#.XbMgFNF7m90. VA cohort was not analyzed for gender specific differences however was noted to be predominately male.
Reporting on race, ethnicity, or other socially relevant groupings	Previously published GWAS study data used as a basis for this work is noted in each study published and is annotated in online GWAS data resources, including UKBB https://www.ebi.ac.uk/gwas/. For LDSC scoring, GWAS data used was available and preprocessed by https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/. For colocalization studies, available summary statistics are provided: https://zenodo.org/record/3518299#.XbMgFNF7m90. In the VA cohort, race, ethnicity, or other social groupings was not annotated.
Population characteristics	Previously published GWAS study data used as a basis for this work is noted in each study published and is annotated in online GWAS data resources, including https://www.ebi.ac.uk/gwas/. For LDSC scoring, GWAS data used was available and preprocessed by https://alkesgroup.broadinstitute.org/LDSCORE/all_sumstats/. For colocalization studies, available summary statistics are provided: https://zenodo.org/record/3518299#.XbMgFNF7m90. VA cohort age was noted to be predominately male and between ages 45-85.
Recruitment	N/A; no persons were recruited for this study.
Ethics oversight	This work was approved by the Stanford Institutional Review Board (IRB) on Human Subjects under the protocol entitled "Public Health Surveillance in the Department of Veterans Affairs." As the project was considered minimal risk, consent to participate was not required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	Sample sizes for each experiment and statistical power calculations are noted in each case in the Supplementary Materials. Sample sizes were chosen to provide sufficient confidence to validate methodological conclusions. Specifically for barcode number. Power analysis for different levels of barcodes at 4 different log2-fold change thresholds was performed fig S1D. We find that sensitivity at our given number of barcodes increases significantly from 32%+/- 17.5% (std) for a fold change of 1.5, to 74% +/- 15.2% for a fold change of 2, and 98% +/- 2.7% for a fold

	change of 3. Thus we demonstrate high sensitivity for high allele-specific signals in our data. Sample size of 2221 variants was chosen based on technical feasibilities during cloning and library preparation, that limited final library size.
Data exclusions	MPRA data for H9-derived D28, HMC3 microglia, and NT2 human embryonal carcinoma cell line were excluded from final analysis due to poor replicability and sequencing quality.
Replication	Replicates are noted with respect to each experiment performed. All experiments were performed with two or four biological replicates and technical duplicates. No data were excluded from the analyses.
Randomization	This work used published GWAS data and retrospective observational clinical data (VA cohort). The experiments were not randomized.
Blinding	This work used published GWAS data and retrospective observational clinical data (VA cohort). No blinding was required. The Investigators were not blinded to allocation during experiments and outcome assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Antibodies

Antibodies used	Normal Rabbit IgG (Cell Signaling Technology #2729s) at 1 ug/ul; rabbit anti-RUNX1 / AML1 antibody (Abcam, #ab23980) at 1 ug/ul; Rabbit antibody H3K27ac (Abcam, ab4729) at 1ug/ul;
Validation	Validation statements were provided on the manufacturer's website. The anti-HA and anti-RUNX1/AML1 were used in Fig. S6E. The Anti H3K27ac is listed as ChIP grade on the manufacturer's website and were used and validated in HiChIP experiments in (Donohue Guo, et al 2022). HiChIP data has been released on GEO, as above.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	ATCC Takara and University of Wisconsin provided these cells with the catalog numbers noted Human ESC line H9 (WA09 line, NIH registry 0046, University of Wisconsin) LentiX HEK293T (Takara, cat. no. 632180) SH-SY5Y neuroblastoma cells (ATCC, CRL-2266) IMR-32 neuroblastoma cells (ATCC, CCL-127) D283 medulloblastoma cells (ATCC, HTB-185) D341 medulloblastoma cells (ATCC, HTB-187)
Authentication	Cell lines were purchased from the manufacturer, source. None of the cell lines were authenticated in lab.
Mycoplasma contamination	All cells tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None were used.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	The VA cohort data is a retrospective analysis and not a randomized control clinical trial.
Study protocol	The serum magnesium study in the VA cohort project was approved by the Stanford University Institutional Review Board under the

Study protocol	protocol entitled "Public Health Surveillance in the Department of Veterans Affairs". As the project was considered minimal risk, consent to participate was not required. Bitscopic is operating under a 10-year Research and Development agreement with the VA signed in 2019.
Data collection	The U.S. Department of Veterans Affairs (VA) healthcare system serves over 9 million veterans at over 1,200 Veterans Health Administration sites of care throughout the United States and U.S. territories63. All VA facilities were included in this analysis, and analysis was inspired by prior work64,65. Patients were examined who'd had a serum magnesium level measured between January 1, 2021 and December 31, 2021. No patients were excluded.
Outcomes	Serum magnesium levels were confirmed in multiple ways to confirm they were a valid test, including using LOINC codes, and ensured that they were measured in the correct unit (mg/dL). Eight neuropsychiatric conditions were identified in the patients using one year of prior ICD-10 codes that were normalized appropriately. These were Alzheimer's Disease, ADHD, Bipolar Disorder, Generalized Anxiety Disorder, Major Depressive Disorder, Obsessive Compulsive Disorder, Parkinson's Disease and Schizophrenia. Chronic kidney disease was included as a control with well documented relationship with magnesium levels. Relative disease prevalence for serum magnesium levels in the bottom 10th and upper 10th deciles, as well as bottom and upper of six quantiles. Given that Alcohol Use Disorder has a well-known effect on Serum Magnesium levels, patients with this condition were identified for subsequent additional analysis to remove a potential confounder. Significance was determined by linear regression.