Mapping the genetic regulation of long non-coding RNAs across cellular contexts advances understanding of complex brain diseases

Jia et al

Supplemental information

Supplemental Note
Supplementary Figures S1-S15

Supplemental Note

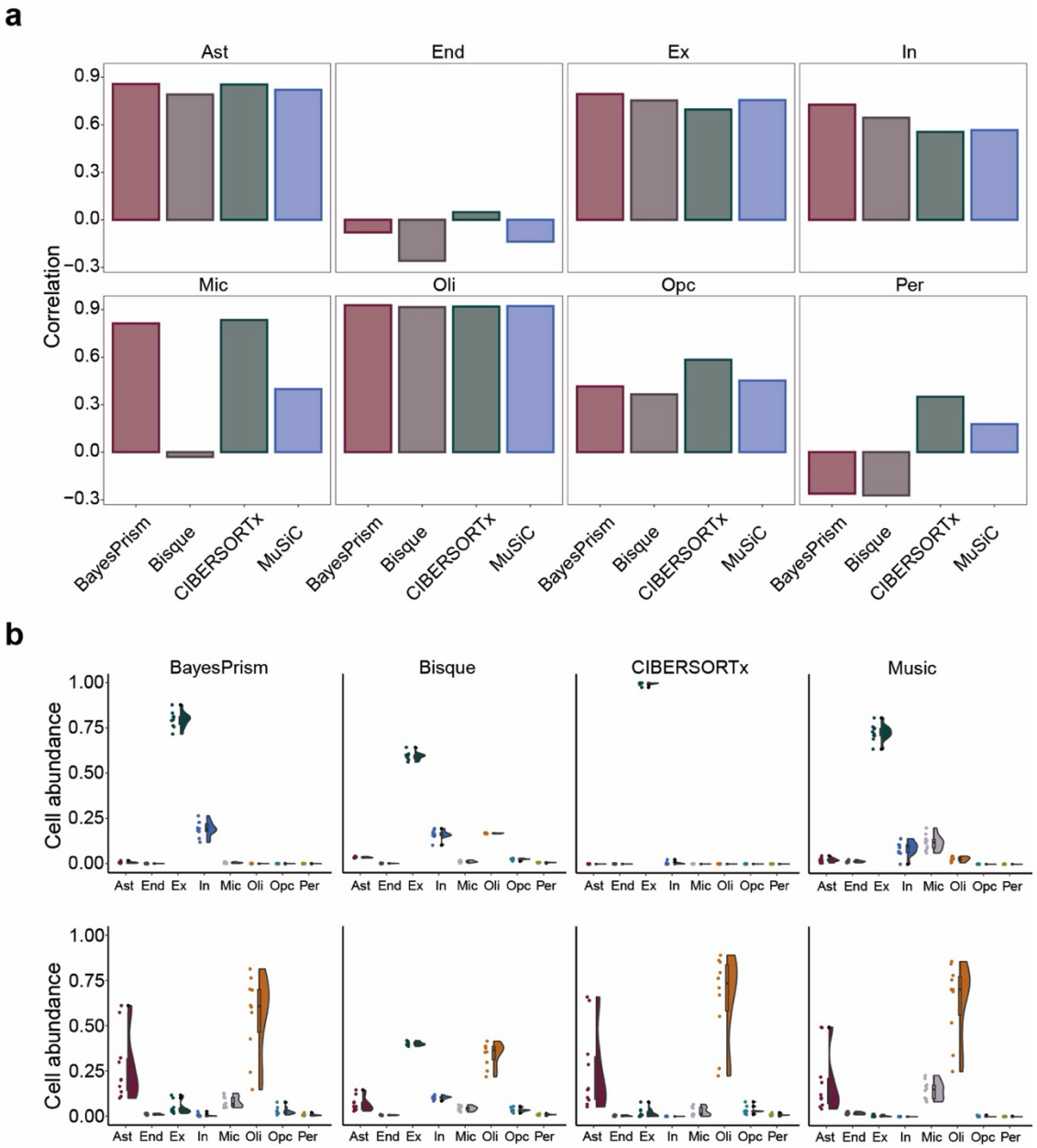
Acknowledgments

HRS (dbGaP accession: phs000428): HRS is supported by the National Institute on Aging (NIA U01AG009740). The genotyping was funded separately by the National Institute on Aging (RC2 AG036495, RC4 AG039029). Genotyping was conducted by the NIH Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotyping quality control and final preparation of the data were performed by the Genetics Coordinating Center at the University of Washington.

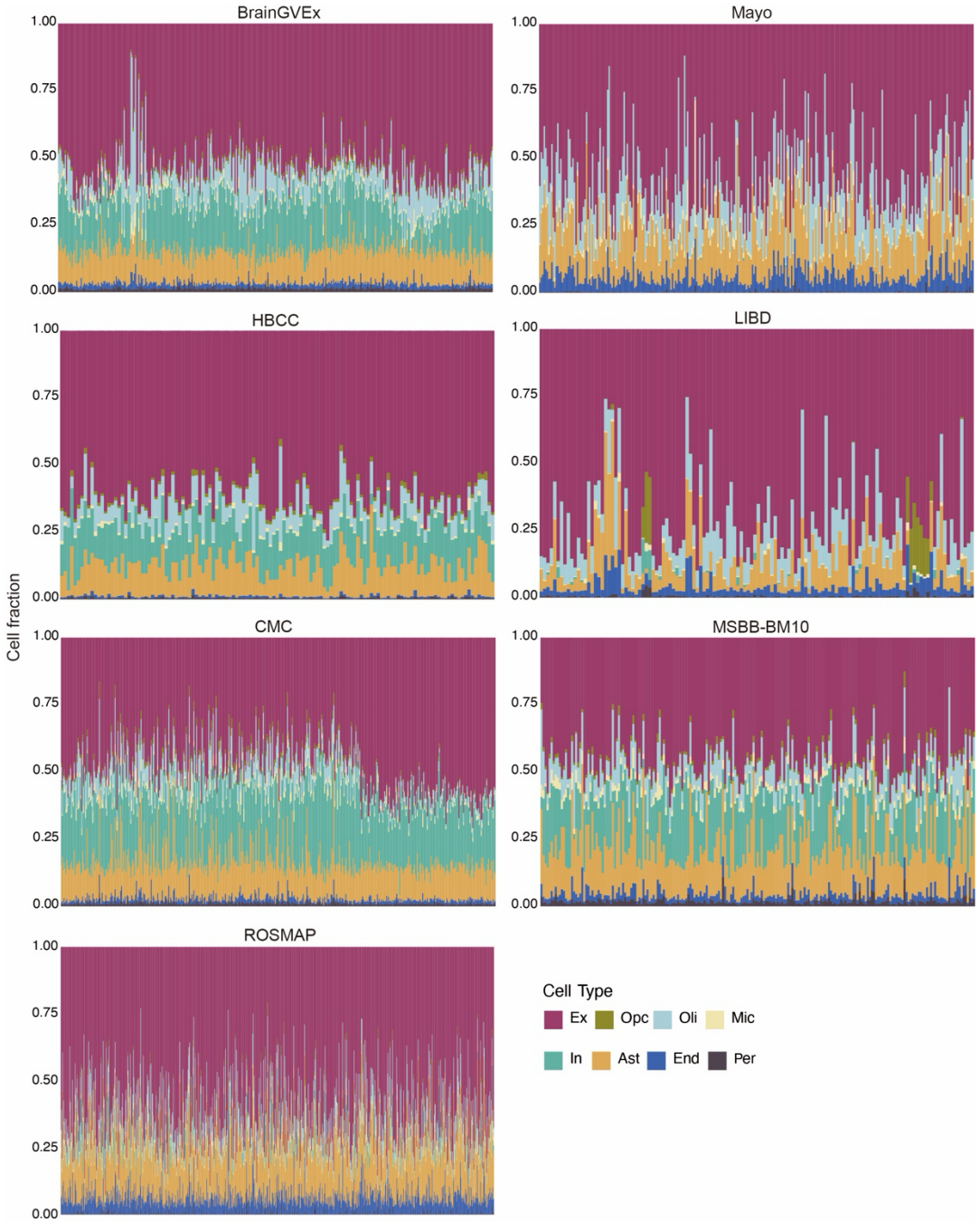
GTEx (dbGaP accession: phs000424): The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\Leidos Biomedical Research, Inc. subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the Broad Institute, Inc. Biorepository operations were funded through a Leidos Biomedical Research, Inc. subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN261200800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101825, & MH101820), the University of North Carolina - Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822).

PsychENCODE (Synapse accession: syn4921369): Data were generated as part of the PsychENCODE Consortium supported by U01MH103339, U01MH103365, U01MH103392, U01MH103340, U01MH103346, R01MH105472, R01MH094714, R01MH105898, R21MH102791, R21MH105881, R21MH103877, and P50MH106934 awarded to Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (USC), Mark Gerstein (Yale), Daniel Geschwind (UCLA), Thomas M. Hyde (LIBD), Andrew Jaffe (LIBD), James A. Knowles (USC), Chunyu Liu (UIC), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Nenad Sestan (Yale), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (UCSF), Patrick Sullivan (UNC), Flora Vaccarino (Yale), Sherman Weissman (Yale), Kevin White (UChicago) and Peter Zandi (JHU).

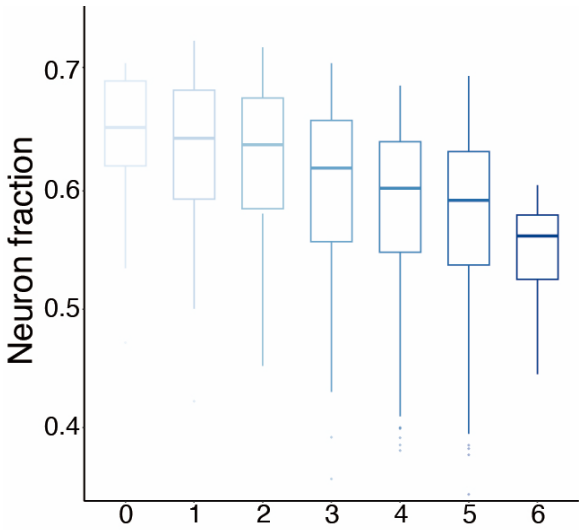
Supplementary Figures



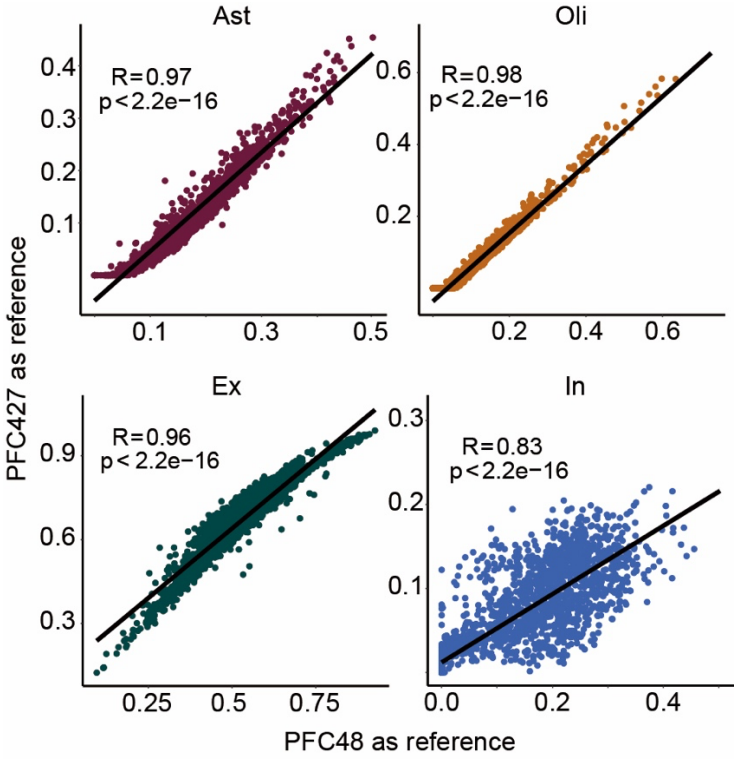
Supplementary Figure S1. Benchmarking cell-type deconvolution accuracy using pseudo-bulk and lineage-sorted data. a. Deconvolution accuracy for eight major brain cell types evaluated on pseudo-bulk samples generated from single-nucleus RNA-seq profiles of prefrontal cortex tissue (PFC427[1], n = 427). Bars show correlations between estimated and true cell-type proportions across four methods. **b.** Deconvolution results for bulk RNA-seq data derived from fluorescence-activated cell sorting (FACS)–purified cortical nuclei with well-defined biological identities from the SuperAgerEpiMap project[2]. The upper panels show deconvolution estimates for $NeuN^+$ (neuronal) nuclear samples, and the lower panels show results for $NeuN^-$ (non-neuronal) nuclear samples, evaluated using BayesPrism[3], MuSiC[4], CIBERSORTx[5], and Bisque[6]. Cell type abbreviations: Ast, astrocytes; End, endothelial cells; Ex, excitatory neurons; In, inhibitory neurons; Mic, microglia; Oli, oligodendrocytes; OPC, oligodendrocyte precursor cells; Per, pericytes.



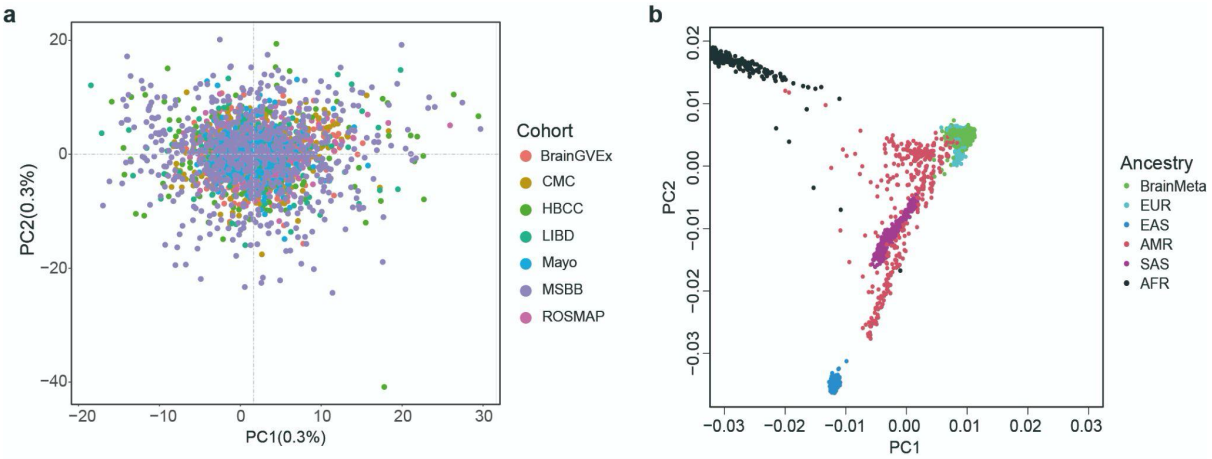
Supplementary Figure S2. Cell-type deconvolution across cortical RNA-seq cohorts using the PFC48[7] as single-cell reference. Bulk cortical RNA-seq data from seven independent cohorts, including BrainGVEx, Mayo, HBCC, LIBD, CMC, MSBB-BM10, and ROSMAP, were deconvolved into eight major brain cell types using BayesPrism, with the PFC48 dataset as the reference. The inferred cell types include excitatory neurons (Ex), inhibitory neurons (In), astrocytes (Ast), oligodendrocytes (Oli), oligodendrocyte precursor cells (OPC), microglia (Mic), endothelial cells (End), and pericytes (Per). Each bar represents one individual sample, with bar height indicating the estimated cellular composition.



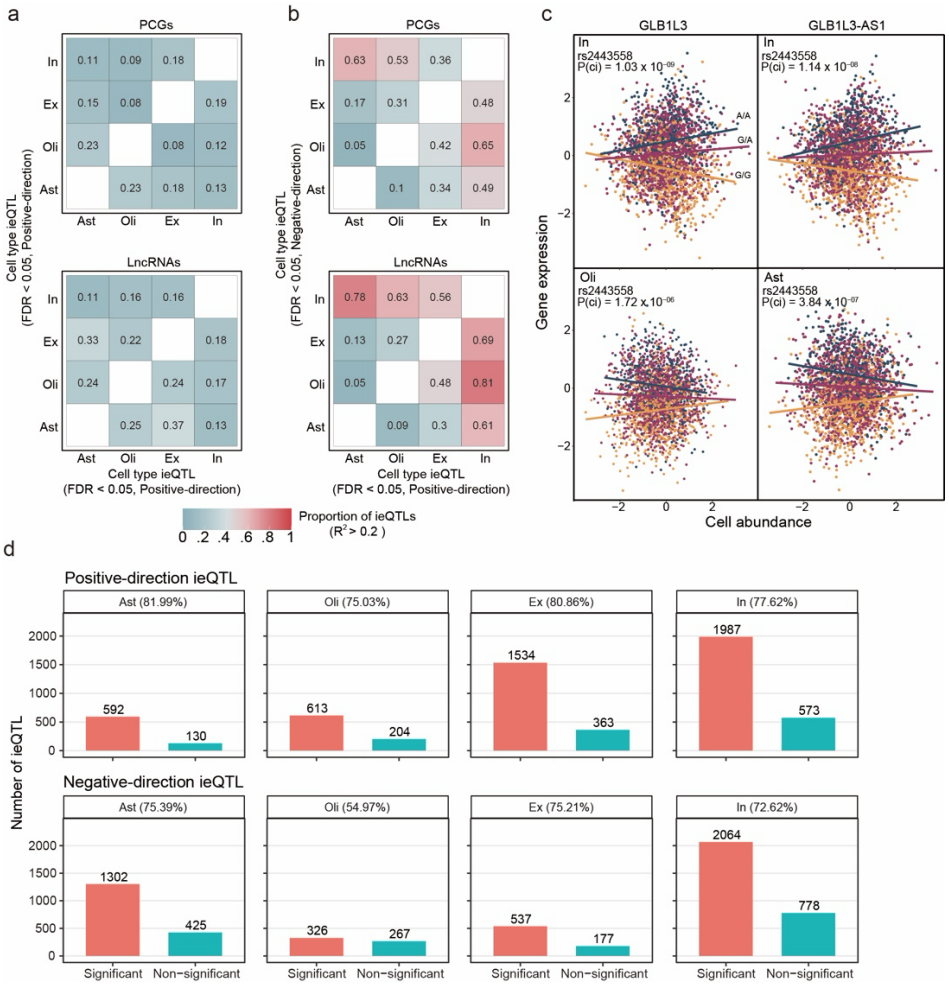
Supplementary Figure S3. Estimated neuronal proportions for ROSMAP cortical RNA-seq samples stratified by Braak stage. Neuronal fractions were inferred for bulk cortical RNA-seq samples from the ROSMAP cohort using BayesPrism, with the PFC48 single-nucleus RNA-seq dataset as the reference. The y-axis shows the estimated neuronal fraction inferred from bulk RNA-seq data, and the x-axis denotes Braak stages (0–6). Braak staging reflects the hierarchical progression of tau pathology in Alzheimer's disease, ranging from stage 0 (no detectable neurofibrillary tangles) to stage VI (widespread neocortical involvement)[8]. Distributions illustrate a stage-dependent decrease in inferred neuronal content with increasing Braak stage.



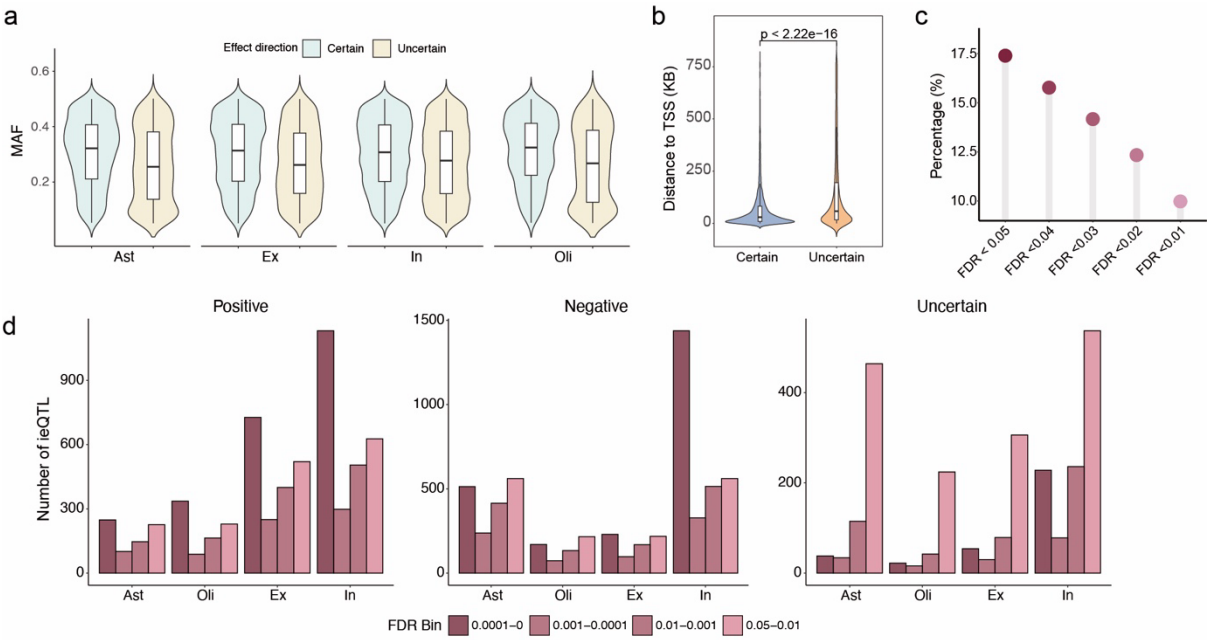
Supplementary Figure S4. Consistency of cell-type deconvolution results across alternative single-cell references. Scatter plots show the correlation between cell-type proportions inferred by BayesPrism using PFC48 (x-axis) and PFC427 (y-axis) as the single-cell reference, respectively. Spearman correlation coefficients (R) and corresponding P-values are shown. Analyses are presented for the four cell types (Ast, Oli, Ex, In) included in ieQTL mapping.



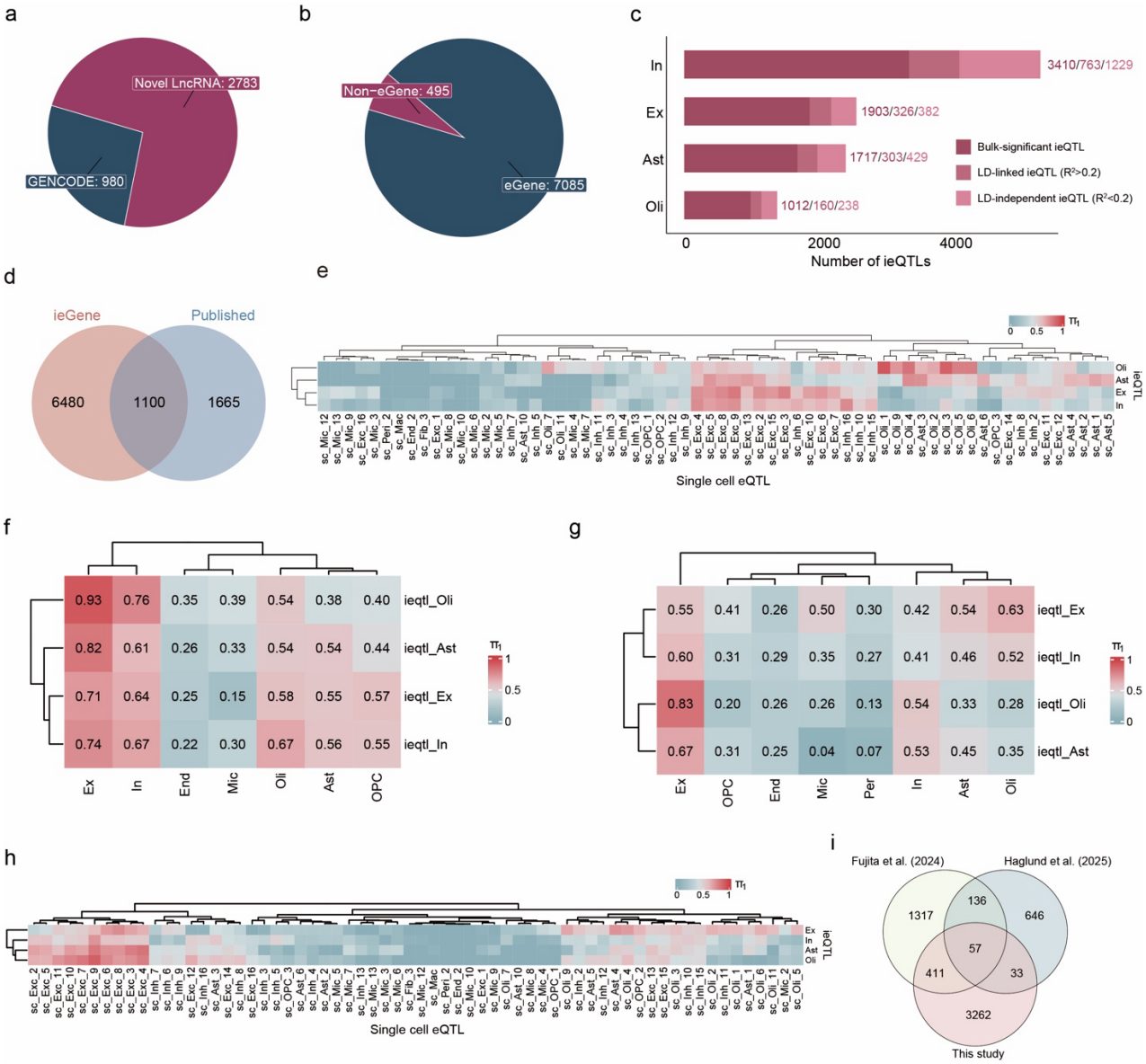
Supplementary Figure S5. Principal component analysis (PCA). Integration of transcriptomic and ancestry data across cohorts. a. PCA of transcriptome-wide gene expression profiles after integrating cortical RNA-seq data across cohorts. Expression data were combined to increase statistical power, and batch effects were removed using a unified linear regression framework prior to downstream analyses. Samples are colored by cohort, showing effective alignment of expression profiles across datasets. **b.** PCA of genetic variation for the integrated cohort (hereafter referred to as BrainMeta) together with reference populations from the 1000 Genomes Project. Samples cluster predominantly with European reference populations, supporting ancestry homogeneity in the analyzed cohort.



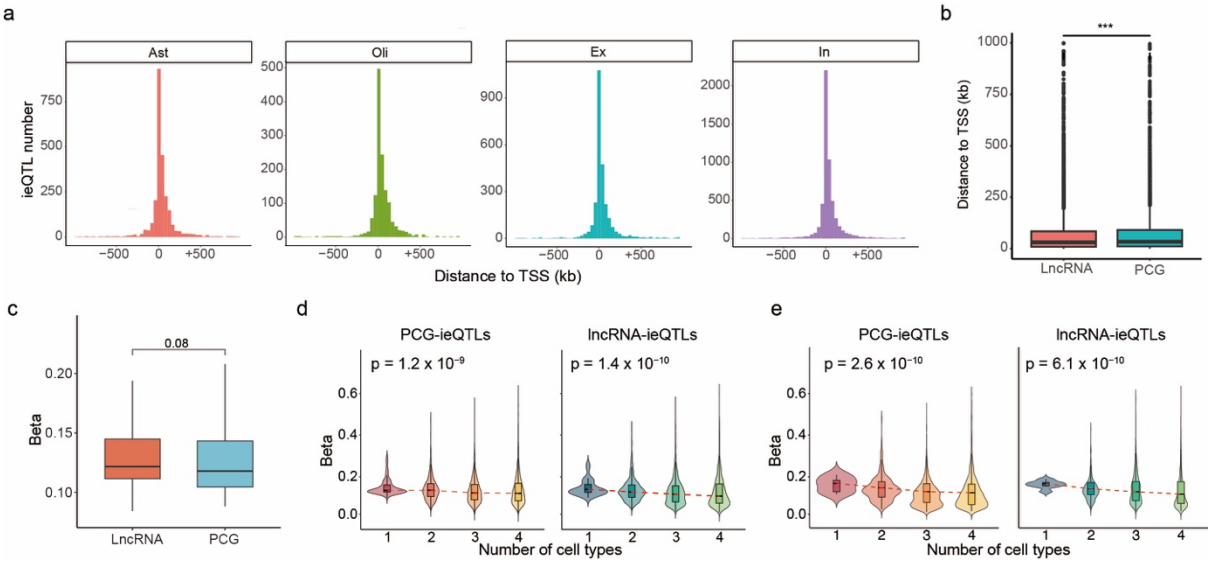
Supplementary Figure S6. Pairwise sharing of cell-type ieQTLs across cell types. a. Pairwise sharing of positive-direction cell-type ieQTLs. b. Pairwise sharing of negative-direction cell-type ieQTLs. In panels a and b, sharing was quantified as the proportion of ieQTLs in linkage disequilibrium (LD, $R^2 > 0.2$) between a query cell type and a validation cell type. For each pairwise comparison, the proportion was calculated using the minimum number of ieQTLs detected in the two cell types as the denominator. The x-axis denotes the query cell type and the y-axis denotes the validation cell type. **c.** An example of negative-direction ieQTL across cell types (rs2443558). The protein-coding gene GLB1L3 and its newly assembled antisense lncRNA GLB1L3-AS1 show a positive direction ieQTL effect in inhibitory neurons, but negative ieQTL effects in oligodendrocytes (Oli) and astrocytes (Ast). Notably, the estimated abundance of inhibitory neurons is inversely correlated with the abundances of oligodendrocytes and astrocytes, providing a potential explanation for the observed opposing interaction effects across cell types. **d.** Differential stability of positive- and negative-direction ieQTLs in joint modeling. Bar plots show the number of ieQTLs that are significant or non-significant in the joint multi–cell-type model across four major brain cell types (Ast, Oli, Ex, In). The top row summarizes positive-direction ieQTLs, while the bottom row summarizes negative-direction ieQTLs. In each panel, bars indicate the number of ieQTLs that remain significant (red) or become non-significant (teal) in the joint model. Percentages shown in parentheses denote the proportion of ieQTLs that remain significant in the joint model for each cell type. Positive-direction ieQTLs consistently exhibit higher joint-model significance rates than negative-direction ieQTLs across all cell types, as supported by Fisher's exact tests comparing significance proportions between directions (Ast, $P = 3.8 \times 10^{-4}$; Ex, $P = 1.7 \times 10^{-3}$; In, $P = 2.5 \times 10^{-5}$; Oli, $P = 5.2 \times 10^{-15}$).



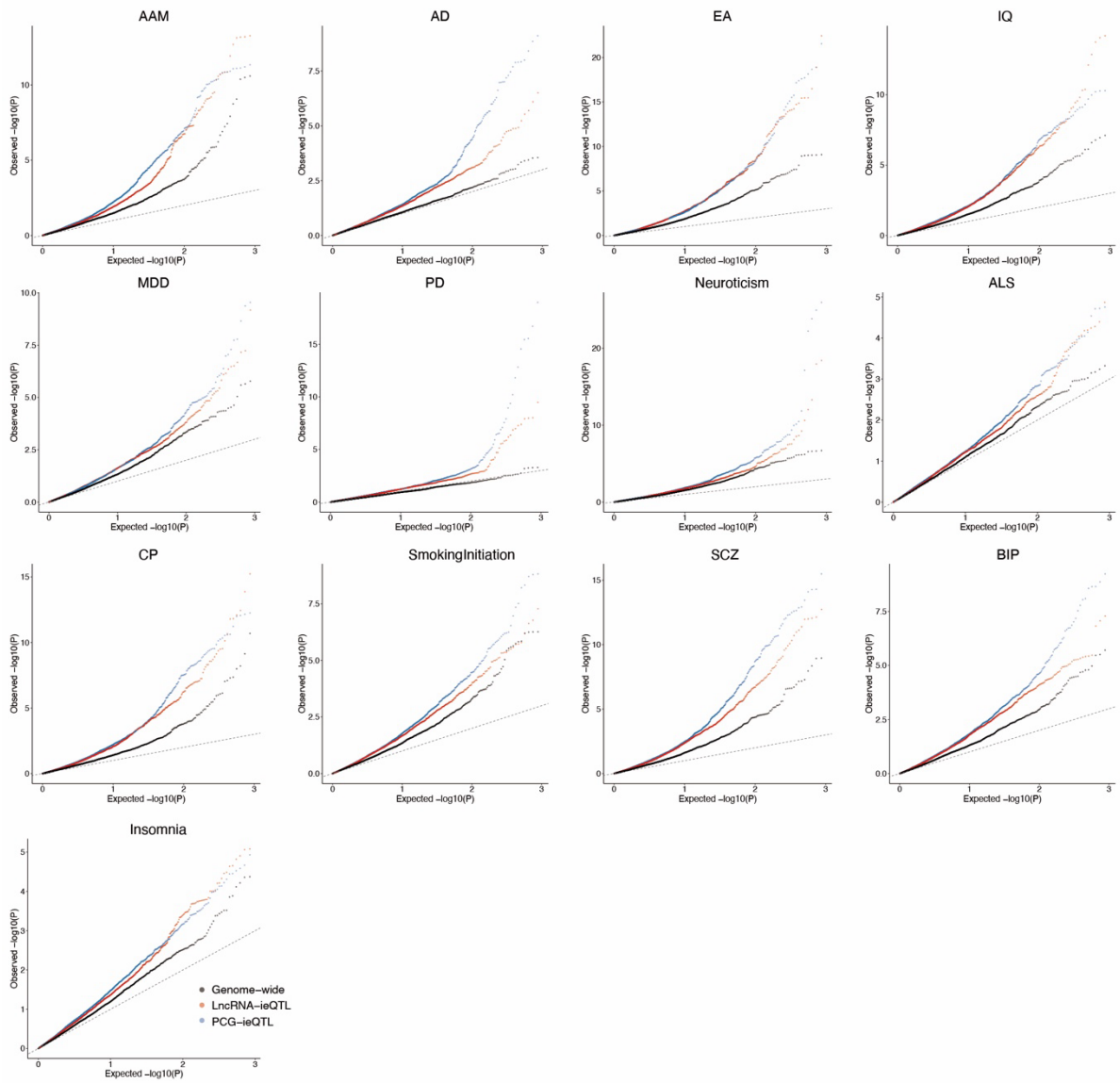
Supplementary Figure S7. Directionality of ieQTL. a. Comparison of effect allele frequency distributions between ieQTLs with well-defined directions (positive or negative) and those classified as directionally ambiguous (uncertain). **b.** Distribution of distances to the transcription start site (TSS) for ieQTLs with defined versus ambiguous directionality. P-values were computed using the Wilcoxon rank test. **c.** As significance thresholds become more stringent, the fraction of directionally ambiguous ieQTLs among all significant ieQTLs progressively declines. **d.** Distribution of ieQTLs with positive, negative, or uncertain directionality across false discovery rate (FDR) intervals.



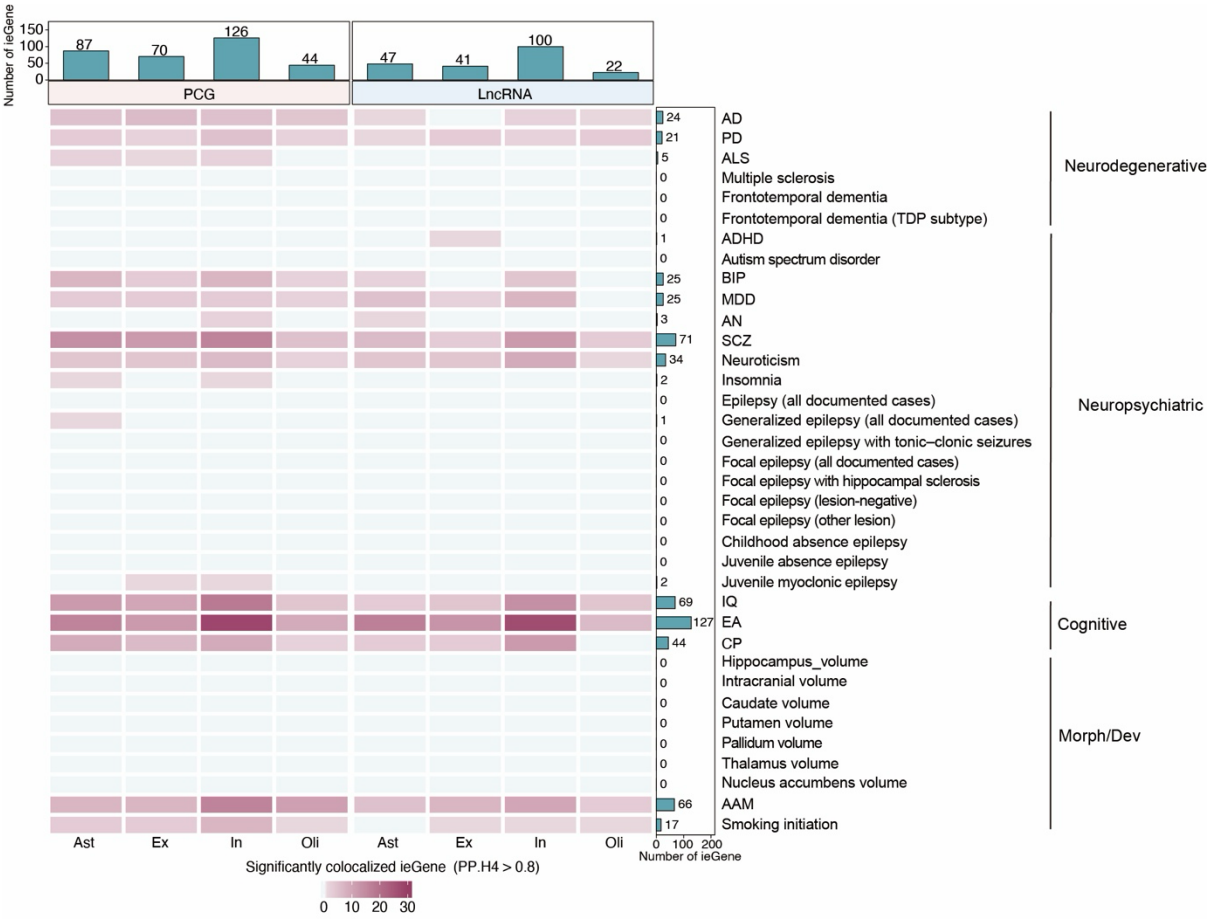
Supplementary Figure S8. Validation of ieQTLs a. Proportion of GENCODE-annotated and newly assembled lncRNAs among all identified ieLncRNAs. **b.** Number of ieGenes identified in this study that harbor at least one cis-eQTL signal in bulk tissue. Among 7,580 ieGenes, 7,085 show evidence of at least one bulk cis-eQTL. **c.** Distribution of ieQTLs across cell types stratified by their linkage disequilibrium (LD) relationship with bulk eQTLs. Bars indicate ieQTLs that coincide with bulk-eQTLs, are in LD with bulk signals ($R^2 > 0.2$), or show low LD with any conditionally independent bulk eQTL ($R^2 < 0.2$). **d.** Overlap of ieGenes identified in this study with those reported by Klein et al.[9] **e.** Replication of positive-direction ieQTLs assessed using π_1 statistics by comparison with cell subtype–specific eQTL summary statistics from Fujita et al.[10]. **f–h.** Replication of negative-direction ieQTLs assessed using π_1 statistics based on comparisons with single-nucleus eQTL resources. Panels f and g show π_1 values comparing negative-direction ieQTLs identified in this study with single-nucleus eQTL summary statistics from Fujita et al. and Haglund et al.[11], respectively. Panel h presents an replication analysis of negative-direction ieQTLs using the cell subtype–specific eQTL dataset from Fujita et al. Heatmaps depict π_1 values across cell-type combinations, with rows and columns clustered using Euclidean distance and complete linkage. **i.** Overlap between ieLncRNAs identified in this study and elncRNAs reported in two single-nucleus eQTL studies (Fujita et al. and Haglund et al.).



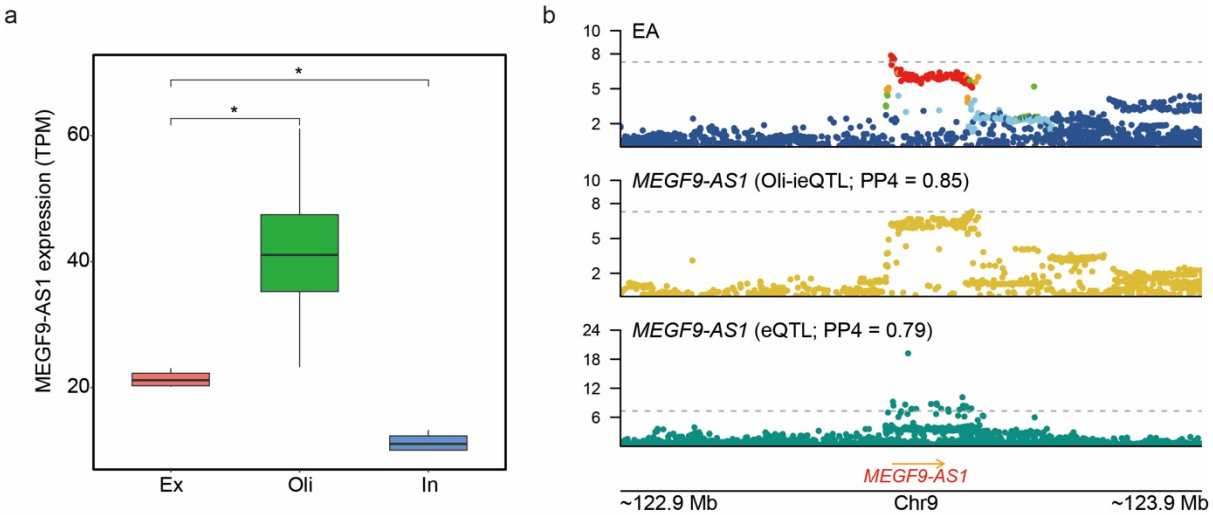
Supplementary Figure S9. Characterization of ieQTLs. a. ieQTLs show enrichment in the vicinity of the transcription start site (TSS) of their target genes. **b.** Distribution of genomic distances between lncRNA-ieQTLs and PCG-ieQTLs relative to the TSS. lncRNA-ieQTLs are located significantly closer to the TSS than PCG-ieQTLs. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. *p < 0.05; **p < 0.01; ***p < 0.001, Wilcoxon signed-rank test. **c.** Boxplots show the distribution of absolute effect sizes of cell-type-specific ieQTLs for lncRNAs and PCGs, estimated from mashr[12] posterior inference. Cell-type-specific ieQTLs were defined as those with a local false sign rate (LFSR) < 0.05 in one and only one cell type. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. **d and e.** Effect sizes of PCG-ieQTLs (left) and lncRNA-ieQTLs (right) as a function of the number of shared cell types. In panel d, ieQTLs were defined using the standard significance threshold (FDR < 0.05), whereas in panel e, the analysis was restricted to ieQTLs passing a more stringent significance threshold ($P_{ieQTL} < 5 \times 10^{-8}$). In both panels, effect sizes decreased significantly with increasing cross–cell-type sharing (Kruskal–Wallis test).



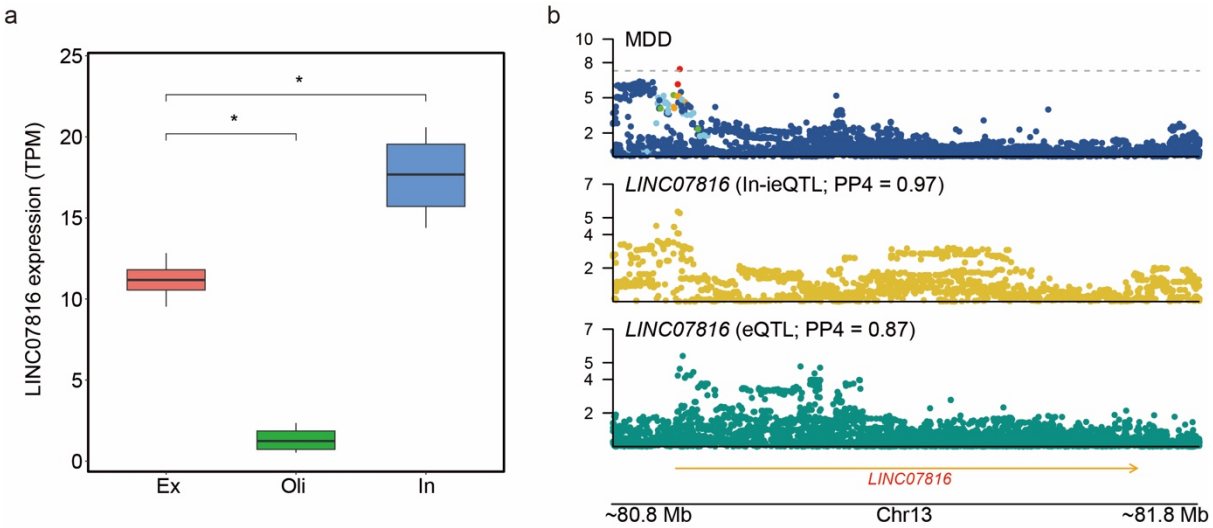
Supplementary Figure S10. Enrichment of lncRNA-ieQTLs in large-scale GWAS. Quantile–quantile (QQ) plots show the distribution of observed versus expected P values for the enrichment of ieQTL-associated variants among GWAS signals from studies with sample sizes greater than 100,000. Each panel represents a distinct trait or disorder. Grey points denote the genome-wide background of all tested SNPs, blue points indicate PCG-ieQTLs, and orange points indicate lncRNA-ieQTLs. Deviation from the null expectation indicates enrichment of GWAS association signals among ieQTLs. Trait abbreviations are as follows: AD, Alzheimer's disease; PD, Parkinson's disease; ALS, amyotrophic lateral sclerosis; MDD, major depressive disorder; BIP, bipolar disorder; SCZ, schizophrenia; IQ, intelligence quotient; EA, educational attainment; AAM, age at menarche; CP, cognitive performance.



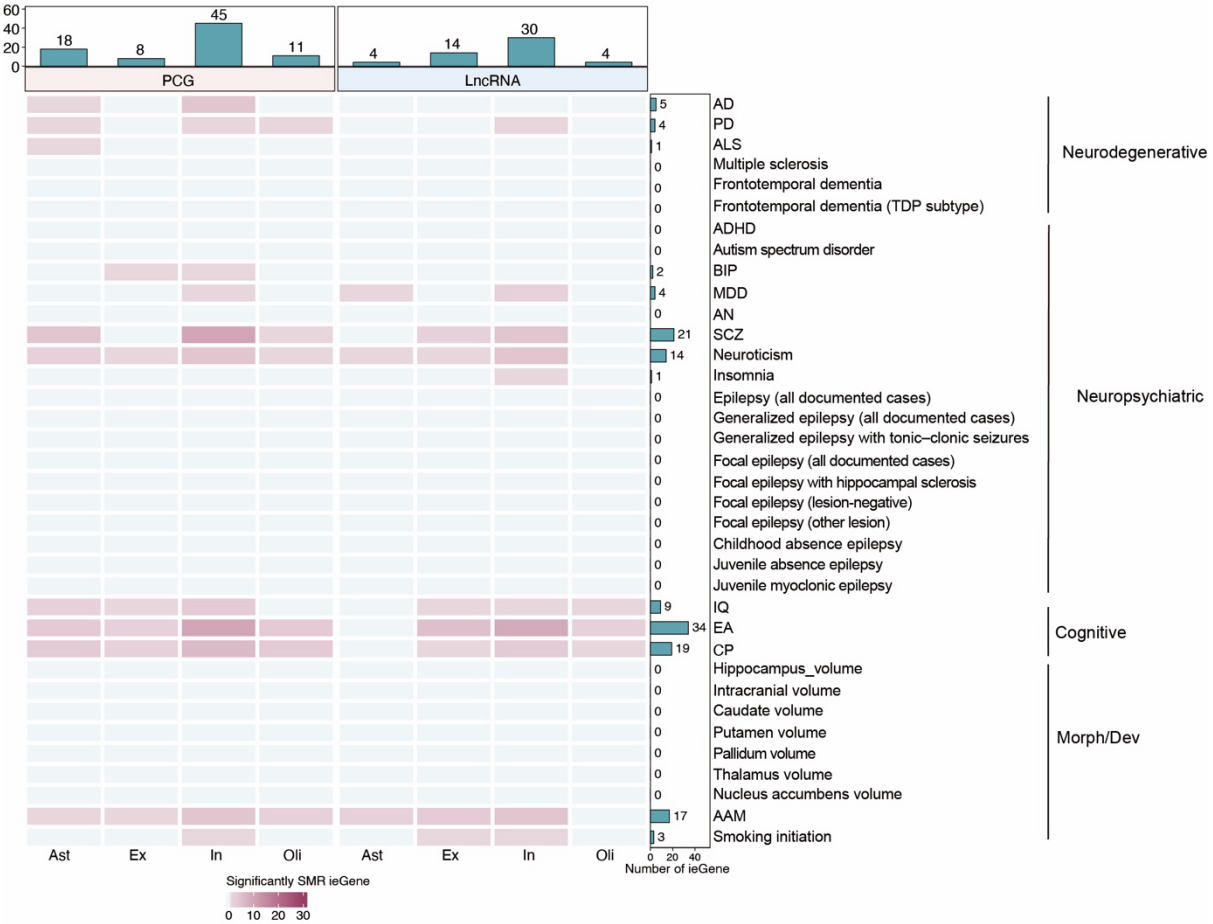
Supplementary Figure S11. Integrative colocalization analysis of ieQTLs with 36 brain-related GWAS traits. Colocalization analysis was performed using COLOC[13] for ieQTLs passing an FDR threshold of < 0.25, with evidence of colocalization defined by a posterior probability of a shared causal variant (PP.H4 > 0.8). Results are shown separately for PCG-ieQTLs and lncRNA-ieQTLs across four brain cell types. For each panel, rows correspond to brain-related traits, grouped into four categories: neuropsychiatric disorders (Neuropsychiatric), neurodegenerative diseases (Neurodegenerative), cognitive traits (Cognitive), and brain morphological or developmental features (Morph/Dev). Columns indicate cell types, and color intensity represents the number of genes with significant colocalization (PP.H4 > 0.8) for a given trait, cell type, and gene category. Bar plots summarize the total number of colocalized genes across cell types for PCGs (left) and lncRNAs (right), while side bars indicate the total number of colocalized genes for each trait across all cell types and gene categories. Trait abbreviations are as follows: AD, Alzheimer's disease; PD, Parkinson's disease; ALS, amyotrophic lateral sclerosis; MDD, major depressive disorder; BIP, bipolar disorder; SCZ, schizophrenia; IQ, intelligence quotient; EA, educational attainment; AAM, age at menarche; CP, cognitive performance; AN, anorexia nervosa.



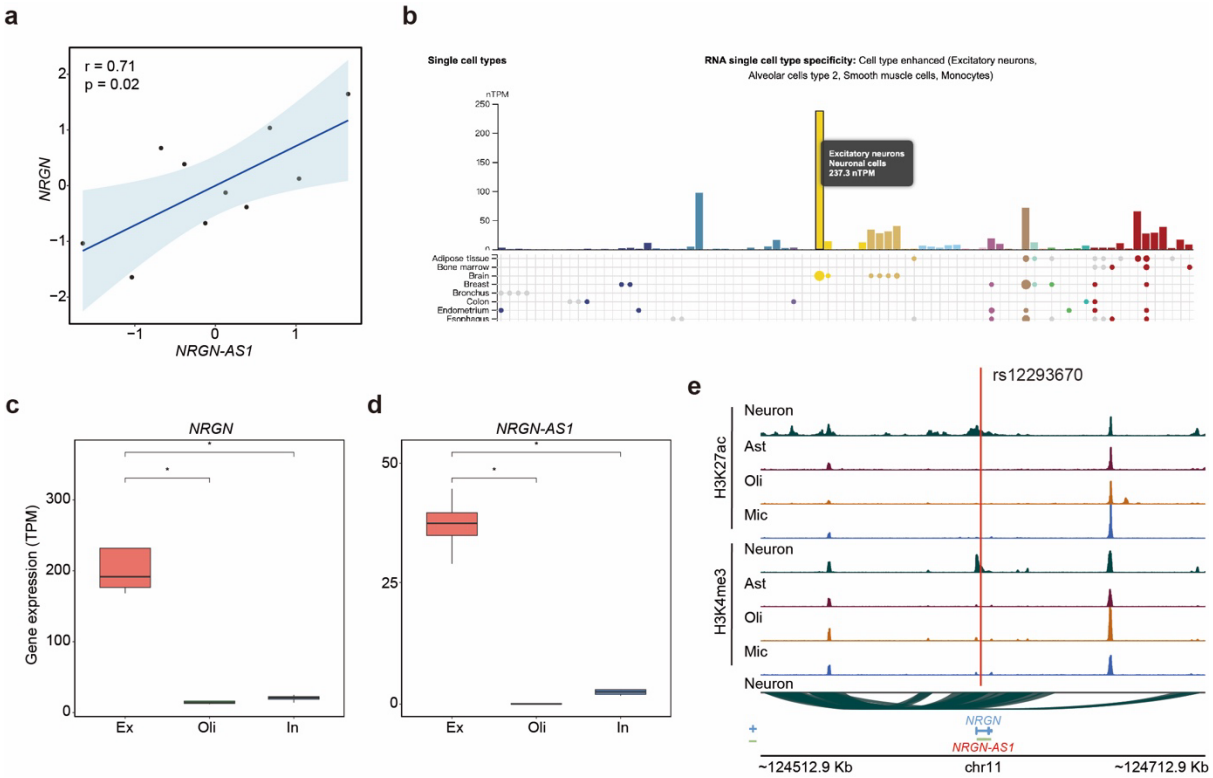
Supplementary Figure S12. Colocalization of *MEGF9-AS1* with EA. a. Expression levels of the newly assembled antisense lncRNA *MEGF9-AS1* in specific cell types: excitatory neurons (Ex), oligodendrocytes (Oli), and inhibitory neurons (In) from EPIGABA samples, highlighting its cell-type-specific expression. **b.** Locus plot showing colocalization of *MEGF9-AS1* in Oli and at the bulk-tissue level with an educational attainment (EA) GWAS signal. Association p-values ($-log_{10}$ scale) are shown for Oli-ieQTLs (middle) and bulk eQTLs (bottom). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. *p < 0.05



Supplementary Figure S13. Colocalization of *LINC07816* with MDD. a. Expression levels of the newly assembled lncRNA *LINC07816* in specific cell types: excitatory neurons (Ex), oligodendrocytes (Oli), and inhibitory neurons (In) from EPIGABA samples, highlighting its cell-type-specific expression. **b.** Locus plot showing colocalization of *LINC07816* in In and at the bulk-tissue level with a major depressive disorder (MDD) GWAS signal. Association p-values ($-log_{10}$ scale) are shown for In-ieQTLs (middle) and bulk eQTLs (bottom). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. *p < 0.05



Supplementary Figure S14. Integrative analysis of ieQTLs with 36 brain-related GWAS traits using SMR. Summary-based Mendelian randomization (SMR) analysis was performed for ieQTLs passing an FDR < 0.25, with significance defined by $p_{SMR} < 0.05/m$ (where m denotes the number of genes tested) and assessed by the HEIDI test ($p_{HEIDI} > 0.01$). Results are shown separately for PCG-ieQTLs and lncRNA-ieQTLs across four brain cell types. Rows correspond to brain-related traits, grouped into four categories: neuropsychiatric disorders, neurodegenerative diseases, cognitive traits, and brain morphological or developmental features. Columns indicate cell types, and color intensity represents the number of genes showing significant SMR associations for a given trait, cell type, and gene category. Bar plots summarize the total number of SMR-significant ieGenes across cell types for PCGs (left) and lncRNAs (right), while side bars indicate the total number of associated ieGenes for each trait across all cell types and gene categories. Trait abbreviations are as follows: AD, Alzheimer's disease; PD, Parkinson's disease; ALS, amyotrophic lateral sclerosis; MDD, major depressive disorder; BIP, bipolar disorder; SCZ, schizophrenia; IQ, intelligence quotient; EA, educational attainment; AAM, age at menarche; CP, cognitive performance; AN, anorexia nervosa.



Supplementary Figure S15. Molecular and epigenomic characterization of *NRGN-AS1* and *NRGN*. a. Pearson correlation between the expression of the newly assembled antisense lncRNA *NRGN-AS1* and *NRGN* in neuronal samples ($NeuN^+$). **b.** Excitatory neuron–specific expression of *NRGN* in the human brain, as reported by the Human Protein Atlas. (https://www.proteinatlas.org). **c–d.** Cell-type–specific expression of *NRGN* (c) and *NRGN-AS1* (d) across excitatory neurons (Ex), inhibitory neurons (In), and oligodendrocytes (Oli). Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. *p < 0.05. **e.** Epigenomic landscape spanning the *NRGN-AS1–NRGN* locus. Tracks show cell-type–specific H3K27ac (top four tracks) and H3K4me3 (next four tracks) ChIP–seq signals in neurons, astrocytes, oligodendrocytes, and microglia, followed by neuronal PLAC-seq interaction profiles and gene annotations.

Supplementary References

1. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365-4385.e27 (2023).
2. Dong, P. *et al.* Population-level variation in enhancer expression identifies disease mechanisms in the human brain. *Nat Genet* **54**, 1493–1503 (2022).
3. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer* **3**, 505–517 (2022).
4. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* **10**, 380 (2019).
5. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773–782 (2019).
6. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* **11**, 1971 (2020).
7. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
8. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol* **82**, 239–259 (1991).
9. de Klein, N. *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet* **55**, 377–388 (2023).
10. Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat Genet* **56**, 605–614 (2024).
11. Haglund, A. *et al.* Cell state-dependent allelic effects and contextual Mendelian randomization analysis for human brain phenotypes. *Nat Genet* **57**, 358–368 (2025).
12. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187–195 (2019).

13. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).