Mapping genetic regulation of gene expression to cellular contexts identifies long non-coding RNAs associated with brain disorders

Jia et al

Supplemental information

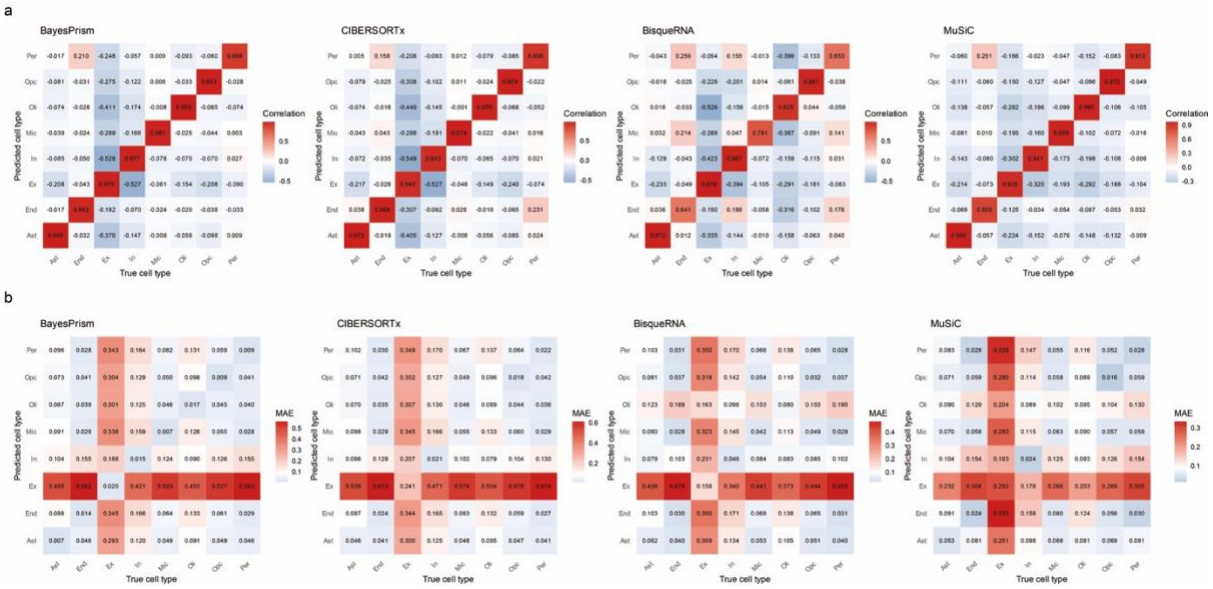
Acknowledgments

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.

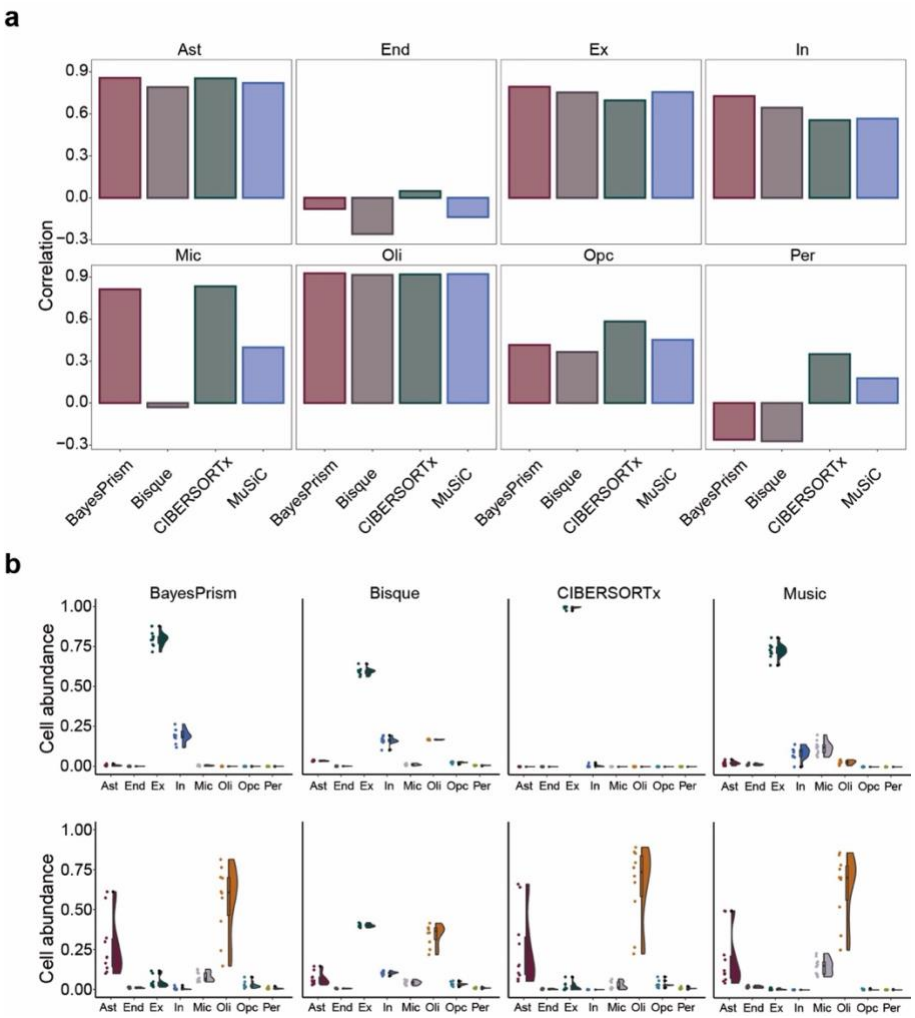
GTEx (dbGaP accession: phs000424): The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled 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).

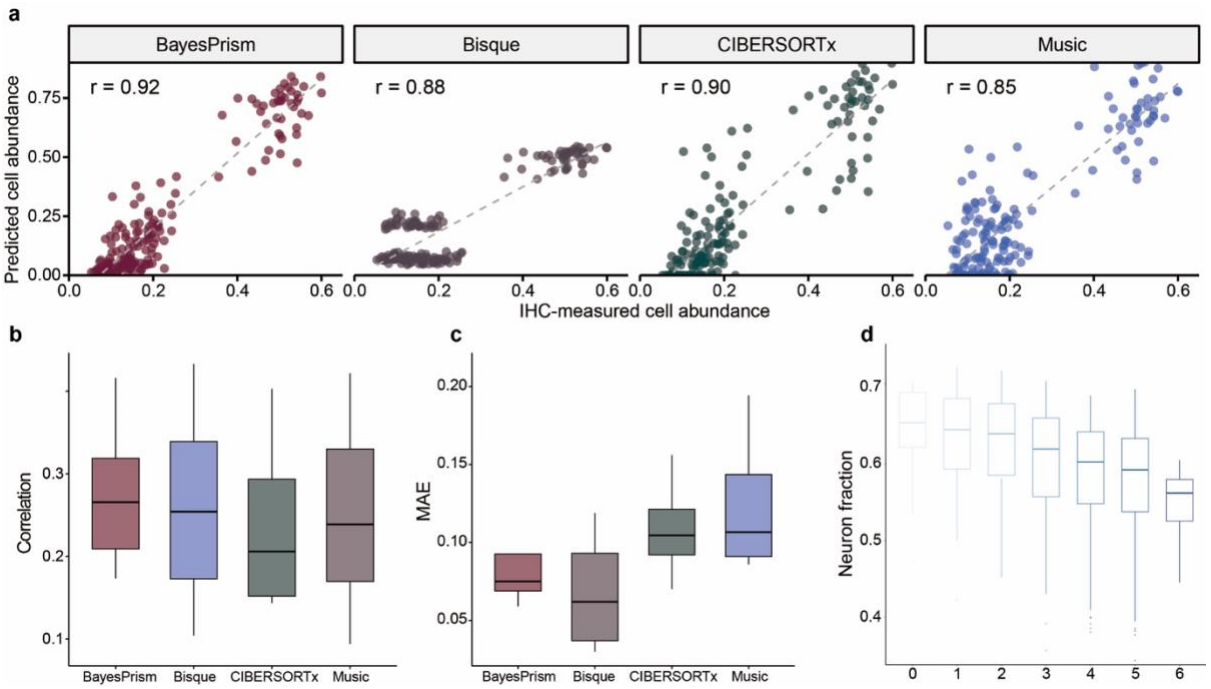
Supplementary Figures



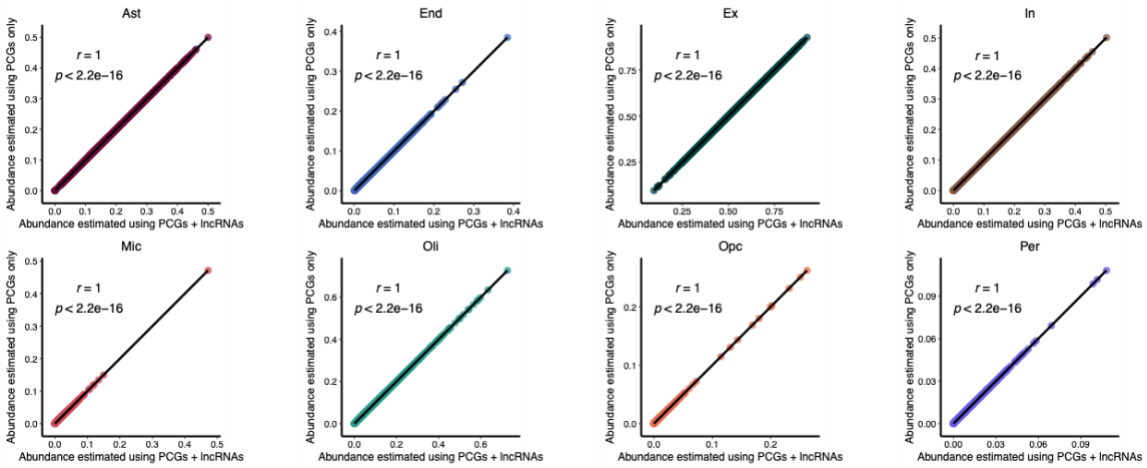
Supplementary Figure 1. Cross-cell-type benchmarking of deconvolution accuracy. a. Spearman correlation heatmaps comparing estimated and reference cell-type proportions for BayesPrism[1], MuSiC[2], CIBERSORTx[3], and Bisque[4]. Rows represent estimated cell types and columns represent reference cell types. A clear diagonal pattern indicates that each estimated cell-type proportion most strongly tracks variation in the corresponding reference cell type. **b.** Cross-cell-type MAE heatmaps comparing estimated and reference cell-type proportions. Colour scales indicate Spearman's rho values in (a) and MAE values in (b). Both heatmaps were computed across 5,000 artificial mixtures (n = 5,000). Abbreviations: Ast, astrocytes; End, endothelial cells; Ex, excitatory neurons; In, inhibitory neurons; Mic, microglia; Oli, oligodendrocytes; Opc, oligodendrocyte precursor cells; Per, pericytes.



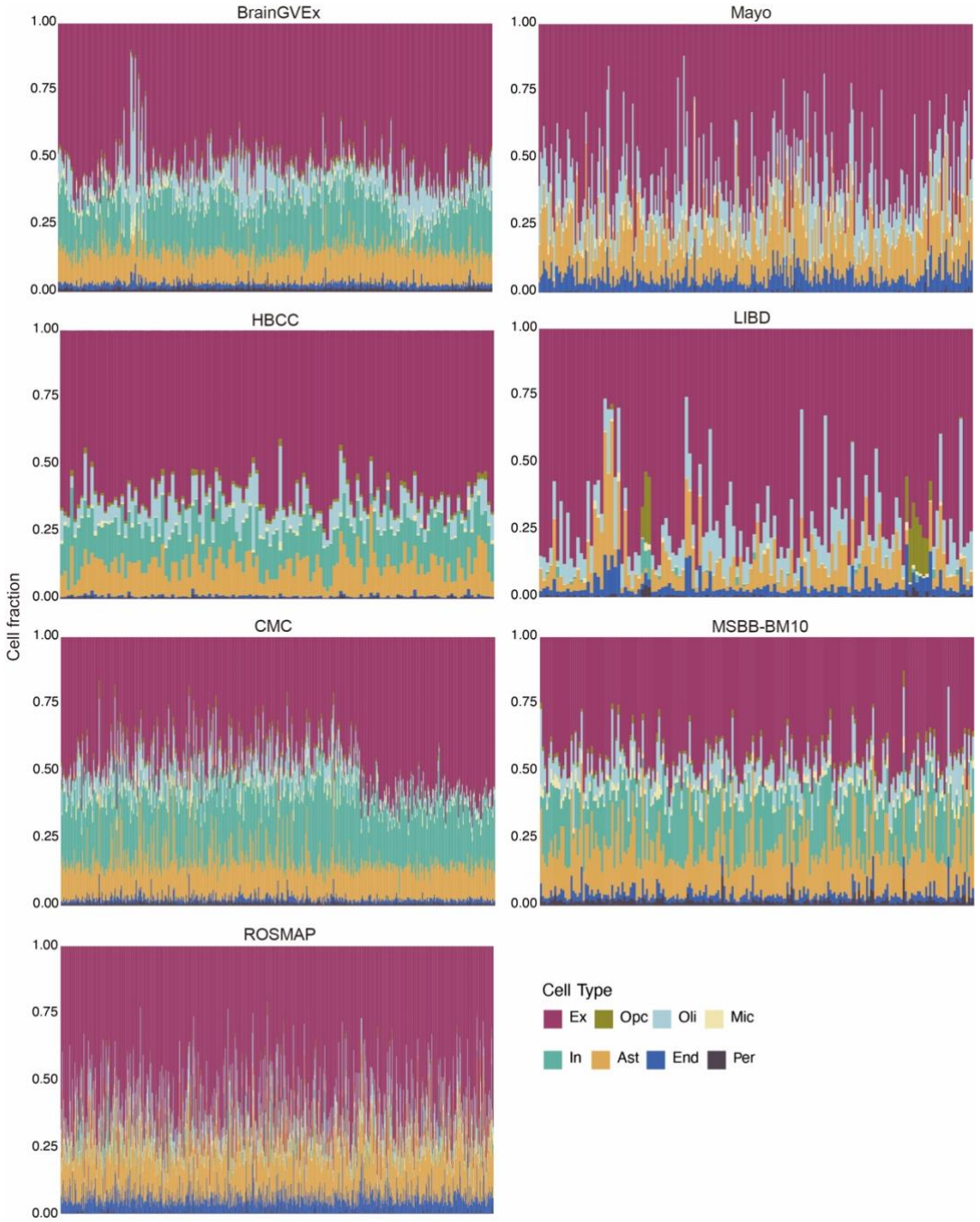
Supplementary Figure 2. Benchmarking cell-type deconvolution accuracy using pseudo-bulk and lineage-sorted data. a. Spearman correlations between estimated and true cell-type proportions for eight major brain cell types, evaluated using pseudo-bulk samples generated from single-nucleus RNA-seq profiles of prefrontal cortex tissue (PFC427[5], n = 427). Bars represent Spearman correlations for the four deconvolution 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[6] (n = 10). 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[1], MuSiC[2], CIBERSORTx[3], and Bisque[4].



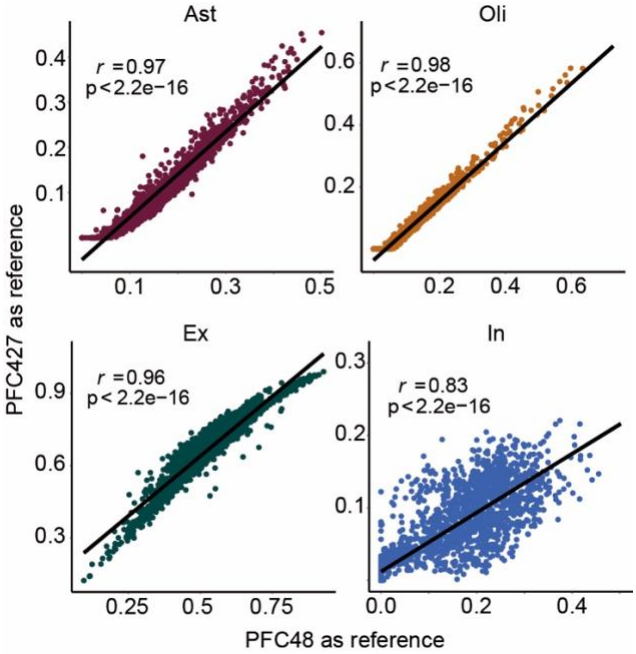
Supplementary Figure 3. Benchmarking and evaluation of deconvolution estimates in ROSMAP cortical samples. a. Spearman correlation between predicted and immunohistochemistry (IHC)-measured cell-type proportions in ROSMAP cortical samples (n = 49; neurons, endothelial cells, microglia, oligodendrocytes and astrocytes)[7]. Each point represents one sample–cell-type pair. **b–c.** Cell-type-level benchmarking of deconvolution performance in ROSMAP cortical samples. Correlation (b) and MAE (c) were calculated separately for each cell type by comparing predicted cell-type proportions with IHC-measured cell-type proportions. IHC provides an independent reference for evaluating deconvolution estimates. However, this comparison should be interpreted with caution because the IHC data are limited in scale and because IHC and bulk RNA-seq reflect different molecular readouts[7–9]. **d.** Estimated neuronal proportions for ROSMAP cortical RNA-seq samples (n = 832) 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)[10]. Distributions illustrate a stage-dependent decrease in inferred neuronal content with increasing Braak stage. For b–d, the box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. Points beyond the whiskers indicate outliers.



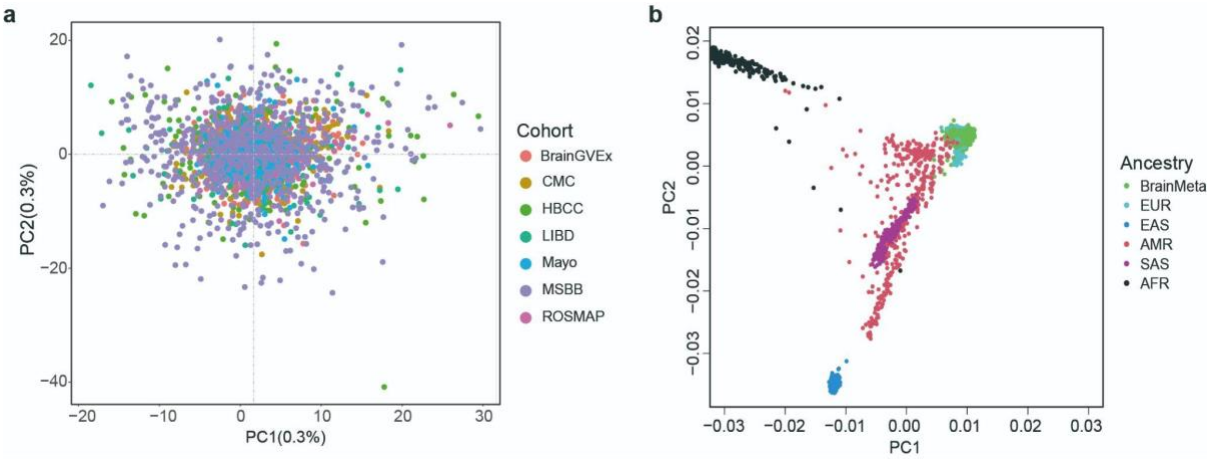
Supplementary Fig. 4. Comparison of BayesPrism-based cell-type abundance estimates using protein-coding genes only and using both protein-coding genes and lncRNAs. Each point represents one cortical sample (n = 2,443). Pearson's correlation coefficients (r) and corresponding two-sided P values are shown for each cell type. The near-identical estimates indicate that inclusion of lncRNAs had negligible impact on deconvolution results.



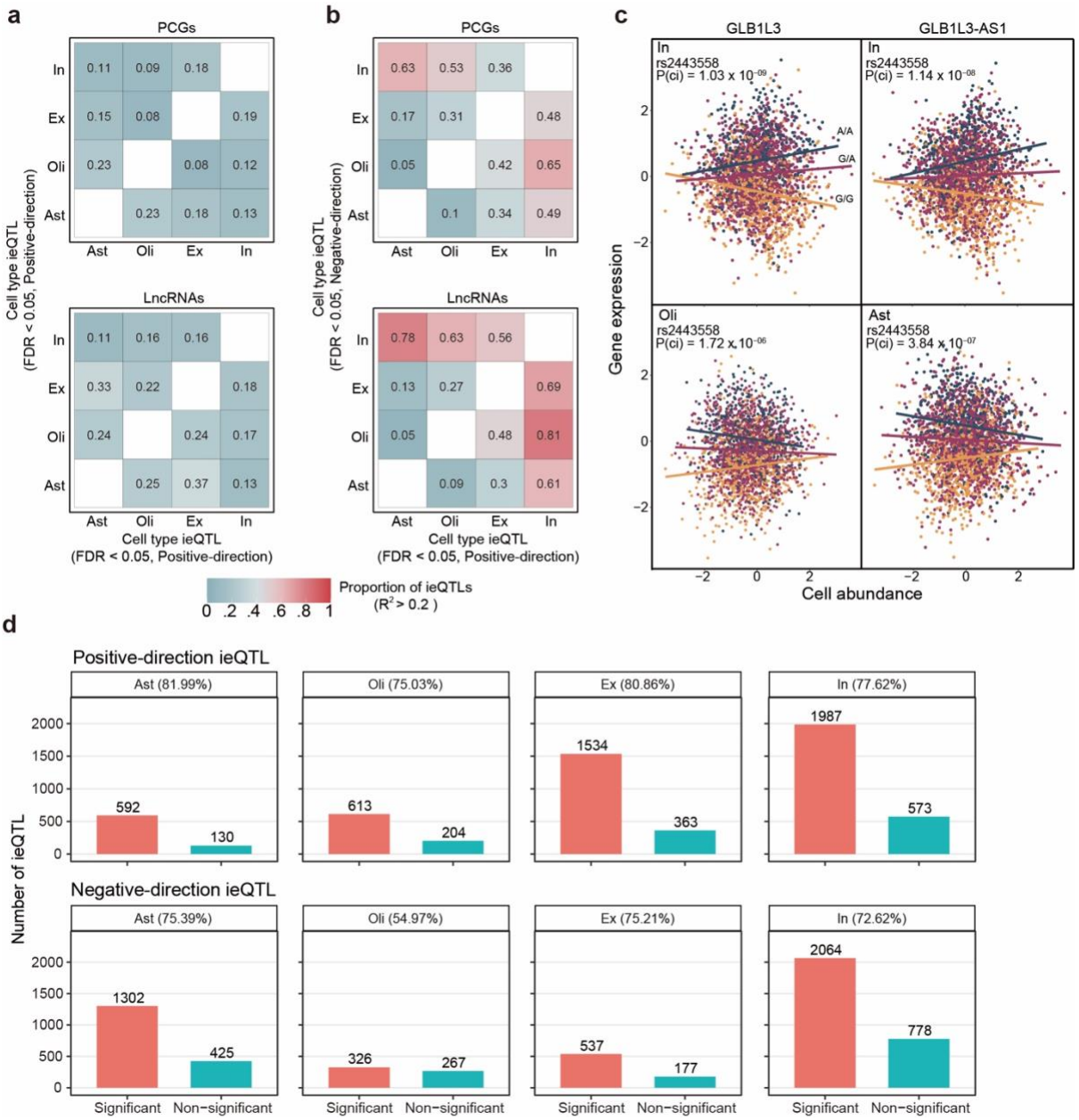
Supplementary Figure 5. Cell-type deconvolution across cortical RNA-seq cohorts using the PFC48[11] as single-cell reference. Bulk cortical RNA-seq data from seven independent cohorts, including BrainGVEx (n = 384), Mayo (n = 288), HBCC (n = 129), LIBD (n = 128), CMC (n = 469), MSBB (n = 213), and ROSMAP (n = 832), were deconvolved into eight major brain cell types using BayesPrism, with the PFC48 dataset as the reference. Each bar represents one individual sample, with bar height indicating the estimated cellular composition.



Supplementary Figure 6. Consistency of cell-type deconvolution results across alternative single-nucleus references. Scatter plots show the correlation between cell-type proportions inferred by BayesPrism using PFC48 (x-axis) and PFC427 (y-axis) as the single-nucleus reference. Each point represents one bulk cortical RNA-seq sample (n = 2,443 samples per cell type). Spearman's ρ and corresponding two-sided P values are shown. No adjustment for multiple comparisons was applied. Analyses are presented for the four cell types (Ast, Oli, Ex and In) included in ieQTL mapping.

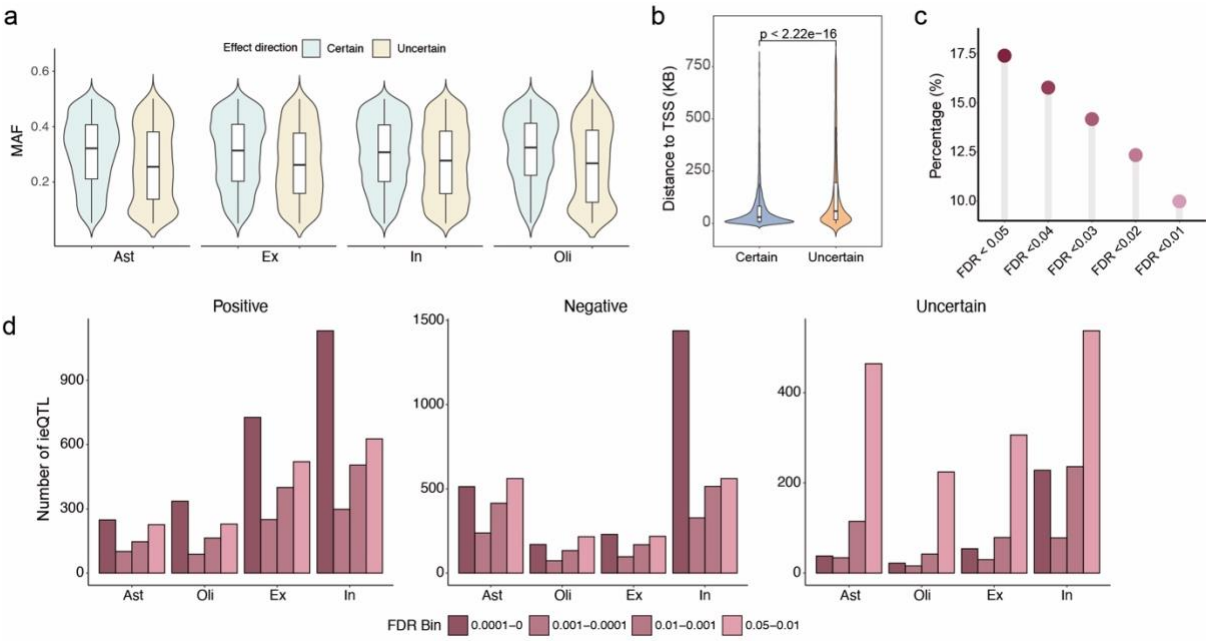


Supplementary Figure 7. Principal component analysis (PCA). Integration of transcriptomic and ancestry data across cohorts. a. PCA of transcriptome-wide gene expression profiles after integrating cortical RNA-seq data across cohorts (n = 2,443). Expression data were combined to increase statistical power, 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; n = 2,443) together with reference populations from the 1000 Genomes Project. Samples cluster predominantly with European reference populations, supporting ancestry homogeneity in the analyzed cohort.

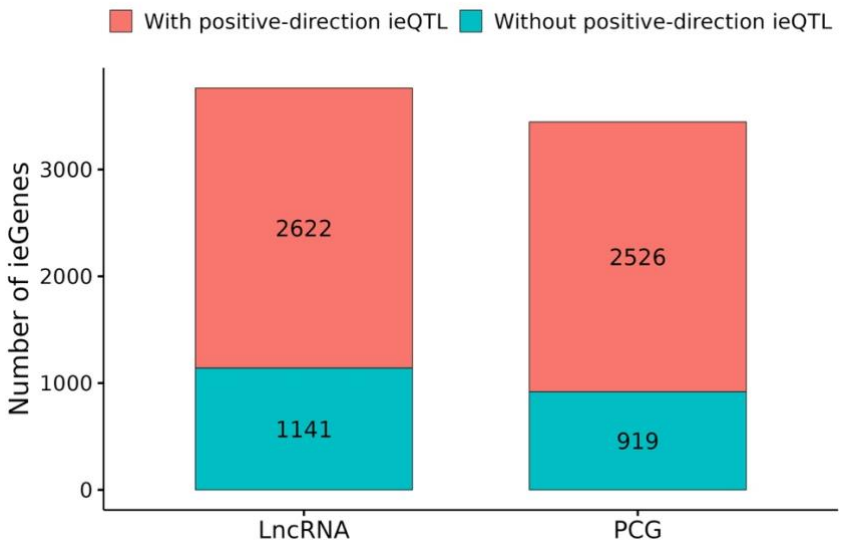


Supplementary Figure 8. Pairwise LD-linked sharing and joint-model stability of ieQTLs. a–b. Pairwise LD-linked sharing of positive-direction (a) and negative-direction (b) ieQTLs across the four analyzed cell types. Sharing was quantified as the proportion of ieQTLs in a query cell type that were LD-linked to ieQTLs detected in a validation cell type, based on LD between their corresponding lead variants ($R^2 > 0.2$). Because marginal ieQTL mapping may identify different but correlated lead variants across cell-type interaction models, an LD-based criterion was used to define sharing. For each pairwise comparison, the sharing proportion was calculated as the number of LD-overlapping ieQTLs divided by the smaller number of ieQTLs detected in the two cell types. The x-axis denotes the query cell type, and the y-axis denotes the validation cell type. **c.** An example of an opposite-direction ieQTL across cell-type interaction models. The ieQTL rs2443558 showed positive interaction effects for GLB1L3 and the unannotated antisense lncRNA GLB1L3-AS1 in inhibitory neurons, but negative interaction effects in Oli and Ast. P values for the interaction effects were obtained from two-sided linear regression tests of the genotype-by-cell-type-proportion interaction term. The estimated abundance of inhibitory neurons was inversely correlated with the abundances of Oli and Ast, highlighting the potential influence of correlated or anticorrelated cell-type abundance

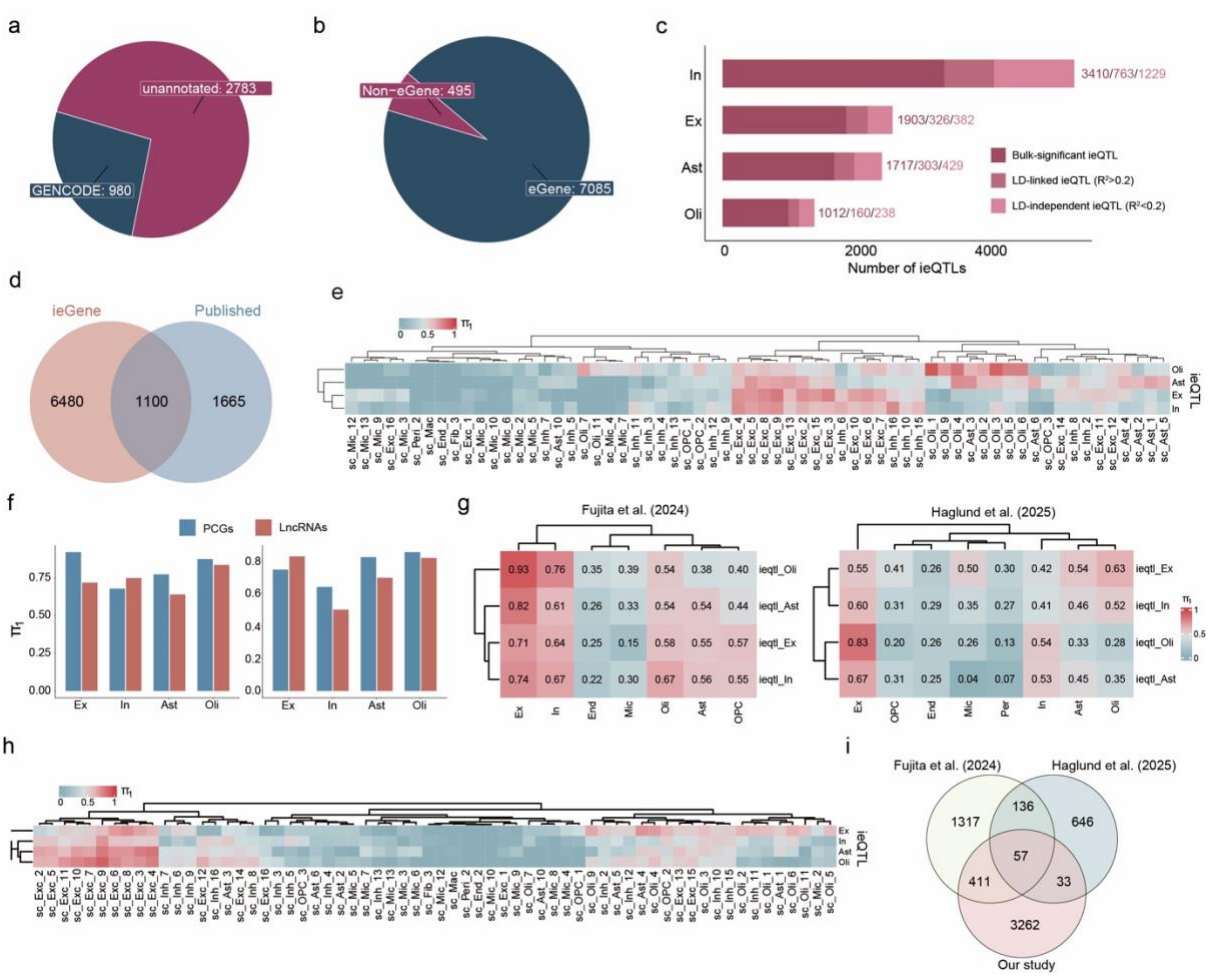
estimates on cell-type interaction models. **d.** Differential stability of positive- and negative-direction ieQTLs in joint modeling. Bar plots show the number of ieQTLs that are significant or non-significant in the joint multi–cell-type model across four major brain cell types (Ast, Oli, Ex, In). 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 retention 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, $P = 5.2 \times 10^{-15}$).



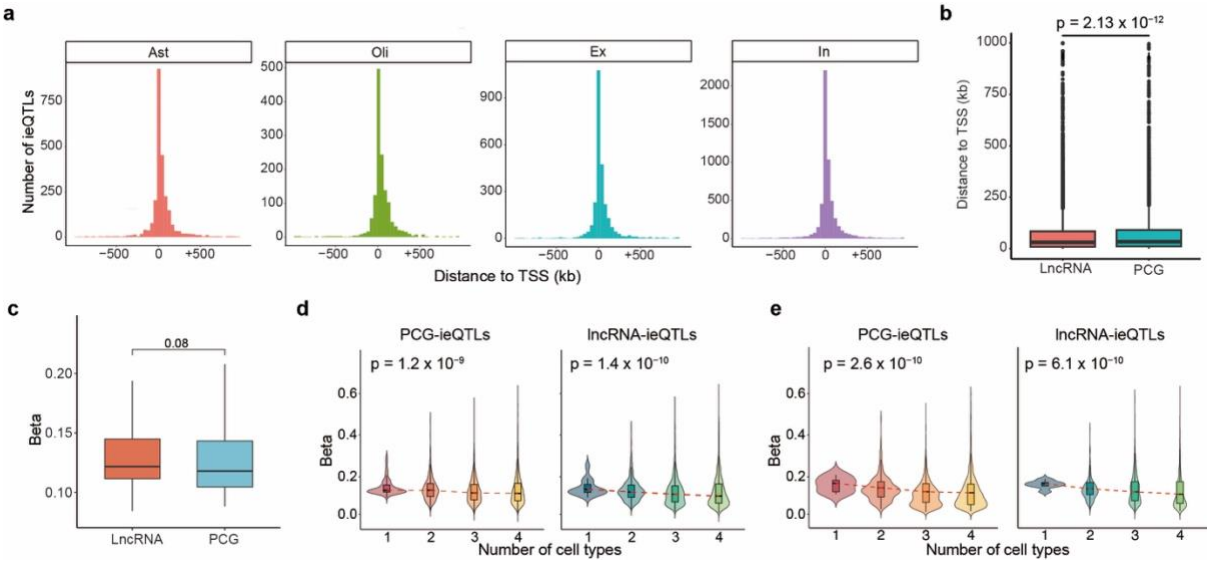
Supplementary Figure 9. Directionality of ieQTL. a. Comparison of minor allele frequency (MAF) distributions between ieQTLs with well-defined directions (positive or negative) and those classified as directionally ambiguous (uncertain). **b.** Distribution of distances to the transcription start site (TSS) for ieQTLs with defined versus ambiguous directionality. P-values were computed using the Wilcoxon rank test. **c.** As the significance threshold becomes 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. For the boxplots in (a) and (b), the box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. Points beyond the whiskers indicate outliers.



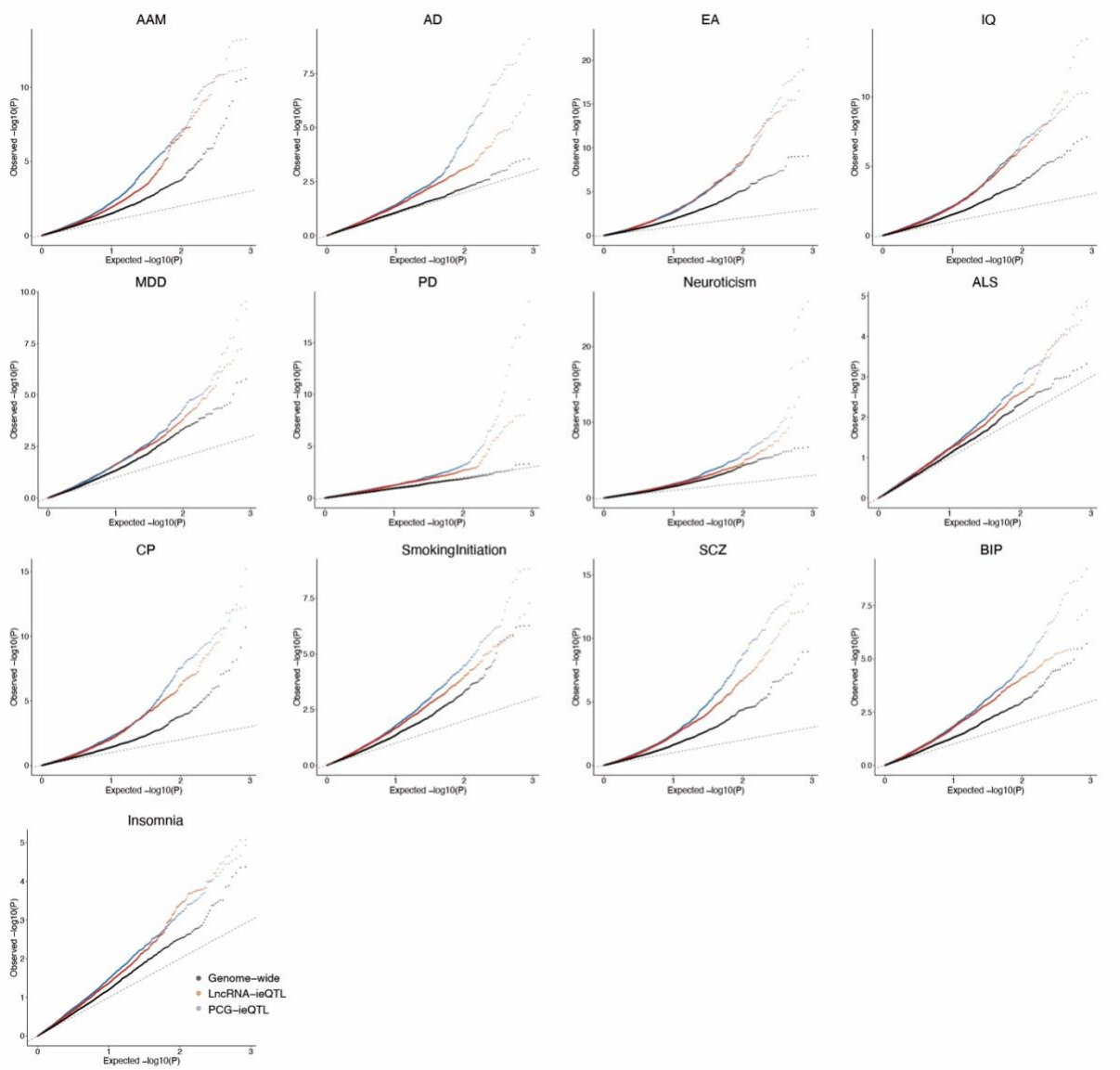
Supplementary Figure 10. Number of ieGenes with or without positive-direction ieQTLs. Stacked bar plot showing the numbers of ieLncRNAs and iePCGs with or without at least one positive-direction ieQTL. Numbers within bars indicate ieGene counts.



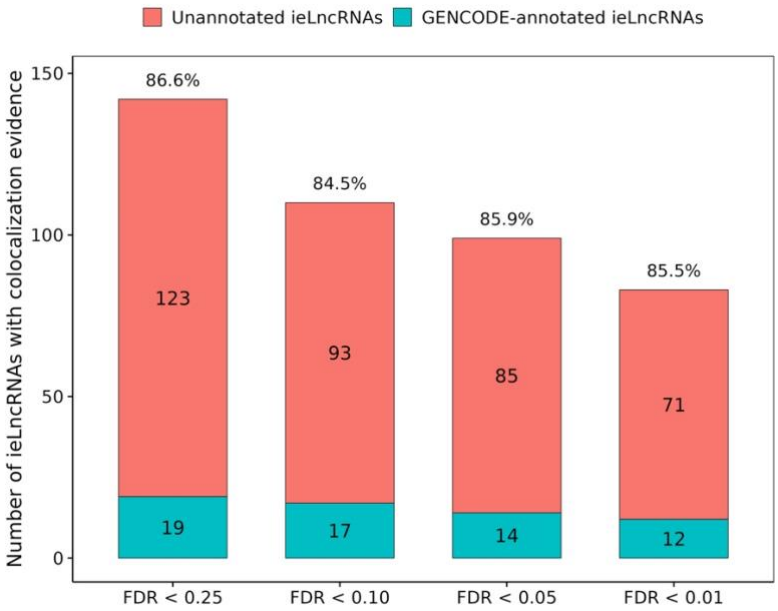
Supplementary Figure 11. Validation of ieQTLs. a. Proportion of GENCODE-annotated and unannotated lncRNAs among all identified ieLncRNAs. **b.** Number of ieGenes identified in this study that harbor at least one cis-eQTL signal in bulk tissue. Among 7,580 ieGenes, 7,085 show evidence of at least one 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.[12]. **e.** Replication of positive-direction ieQTLs assessed using π_1 statistics by comparison with cell subtype–specific eQTL summary statistics from Fujita et al.[13]. **f.** Gene-category-stratified replication of positive-direction ieQTLs. Bar plots show π_1 estimates for PCG-ieQTLs and lncRNA-ieQTLs in matched cellular contexts in the Fujita et al. dataset (left) and the Haglund et al. dataset (right). The greater variability of lncRNA-ieQTL replication may reflect the lower detectability and more limited representation of lncRNAs in current single-cell eQTL datasets. **g–h.** Replication of negative-direction ieQTLs assessed using π_1 statistics based on comparisons with single-nucleus eQTL resources. Panel g shows 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.[14]. Panel h presents a 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.).



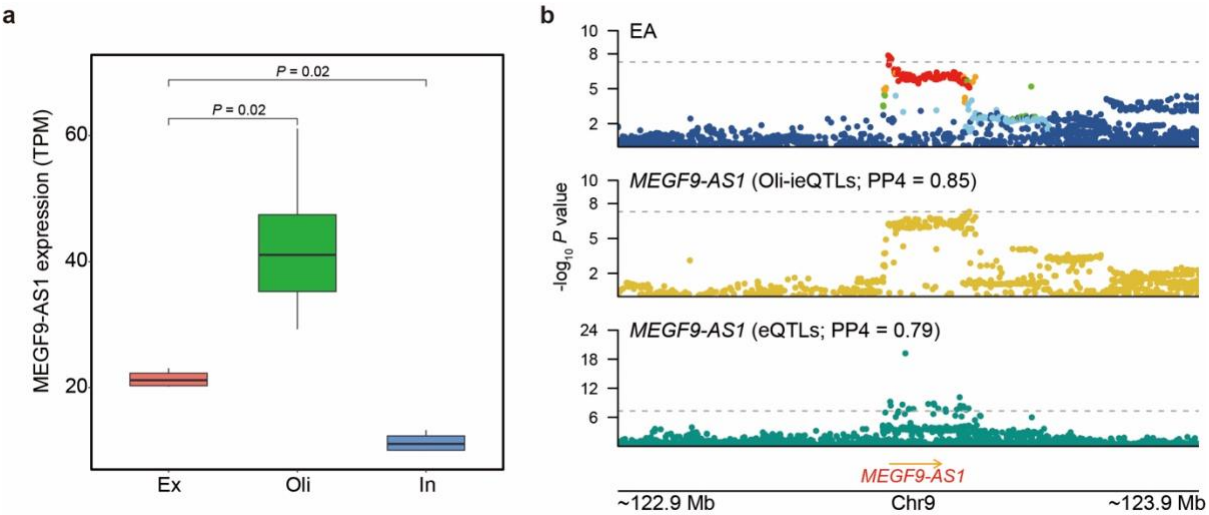
Supplementary Figure 12. Characterization of ieQTLs. a. ieQTLs show enrichment in the vicinity of the transcription start site (TSS) of their target genes. **b.** Distribution of genomic distances between lncRNA-ieQTLs and PCG-ieQTLs relative to the TSS. lncRNA-ieQTLs are located significantly closer to the TSS than PCG-ieQTLs. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. **c.** Boxplots show the distribution of absolute effect sizes of cell-type-specific ieQTLs for lncRNAs and PCGs, estimated from mashr[15] 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–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). For the boxplots in (b–e), the box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. Points beyond the whiskers indicate outliers.



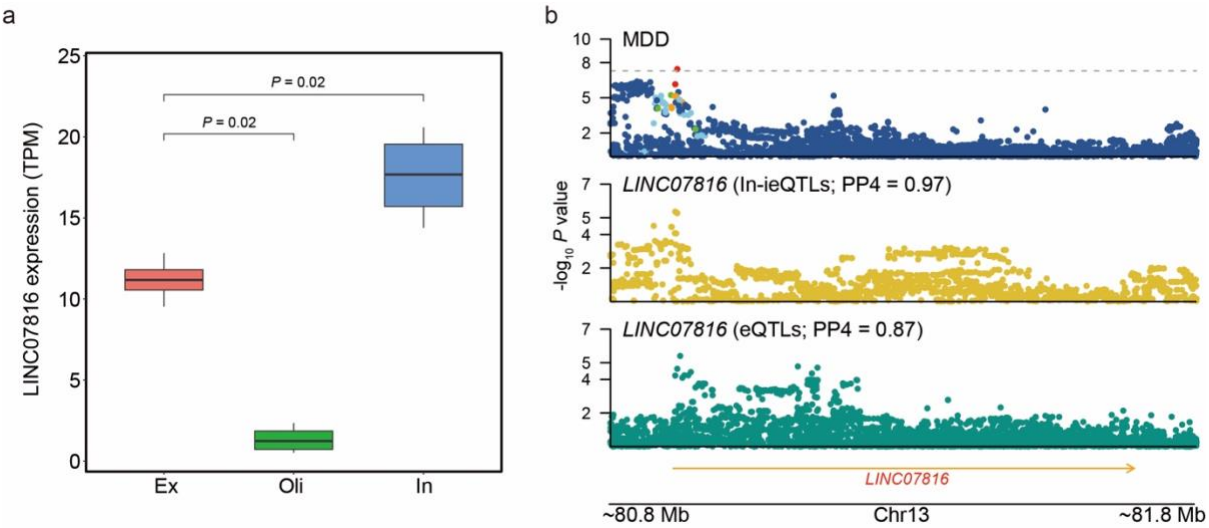
Supplementary Figure 13. Enrichment of lncRNA-ieQTLs in large-scale GWAS. Quantile–quantile (QQ) plots show the distribution of observed versus expected P values for the enrichment of ieQTL-associated variants among GWAS signals from studies with sample sizes greater than 100,000. Each panel represents 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.



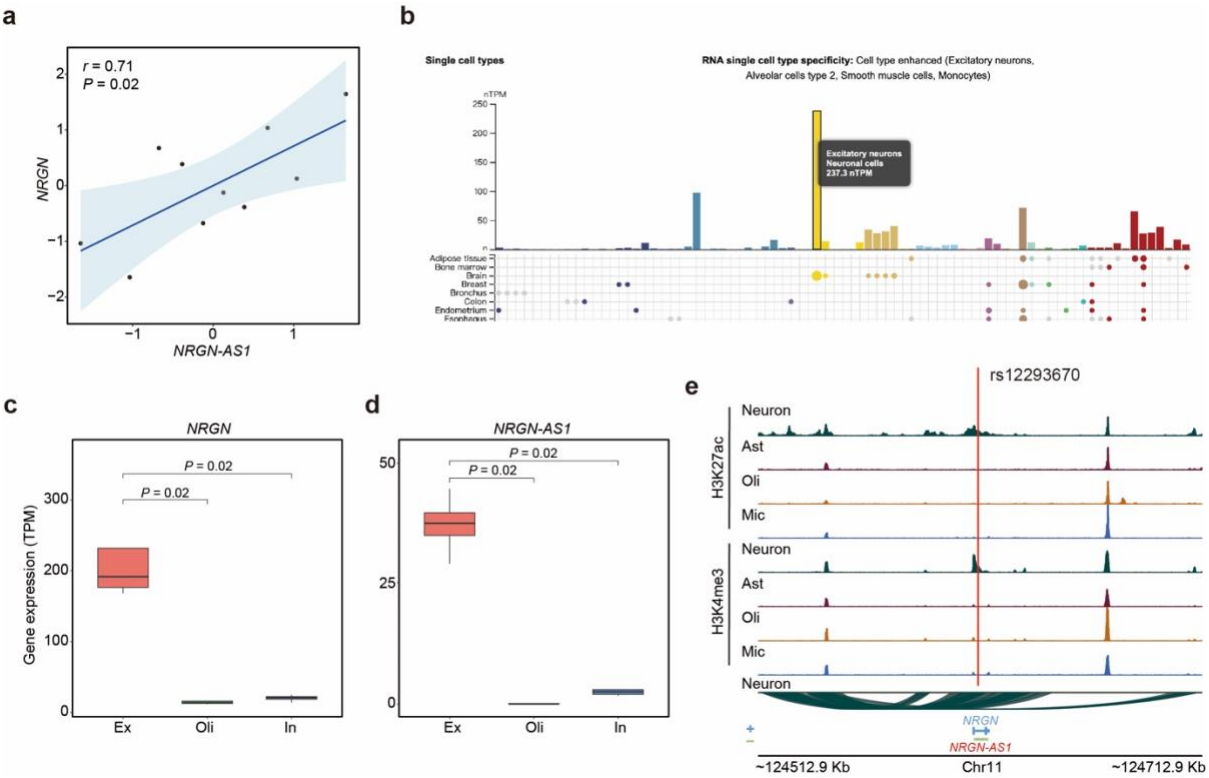
Supplementary Figure 14. Colocalization results across different ieQTL significance thresholds. Stacked bar plot showing the numbers of unannotated and GENCODE-annotated ieLncRNAs with evidence of colocalization at four ieQTL significance thresholds (FDR < 0.25, 0.10, 0.05, and 0.01). Numbers within bars indicate the corresponding numbers of ieLncRNAs, and percentages above bars represent the proportion of unannotated ieLncRNAs among all colocalized ieLncRNAs at each threshold.



Supplementary Figure 15. Colocalization of *MEGF9-AS1* with EA. a. Expression levels of the unannotated antisense lncRNA *MEGF9-AS1* (chr9: 123,348,745–123,437,964) in specific cell types: excitatory neurons (Ex), oligodendrocytes (Oli), and inhibitory neurons (In) from EPIGABA samples (n = 4), highlighting its cell-type-specific expression. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. The box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. **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) were obtained from two-sided linear regression tests of the genotype-by-cell-type-proportion interaction term for Oli-ieQTLs (middle) and the genotype term for bulk eQTLs (bottom).



Supplementary Figure 16. Colocalization of *LINC07816* with MDD. a. Expression levels of the unannotated lncRNA *LINC07816* in specific cell types: excitatory neurons (Ex), oligodendrocytes (Oli), and inhibitory neurons (In) from EPIGABA samples (n = 4), highlighting its cell-type-specific expression. Statistical significance was assessed using a two-sided Wilcoxon rank-sum test. The box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. **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) were obtained from two-sided linear regression tests of the genotype-by-cell-type-proportion interaction term for In-ieQTLs (middle) and the genotype term for bulk eQTLs (bottom).



Supplementary Figure 17. Molecular and epigenomic characterization of *NRGN-AS1* and *NRGN*. a. Pearson correlation between the expression of the unannotated 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 (n = 4). The box indicates the interquartile range (IQR), the line within the box represents the median, and the whiskers extend to the most extreme data points within 1.5 times the IQR. **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. 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).
2. 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).
3. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773–782 (2019).
4. Jew, B. *et al.* Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat Commun* **11**, 1971 (2020).
5. 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).
6. Dong, P. *et al.* Population-level variation in enhancer expression identifies disease mechanisms in the human brain. *Nat Genet* **54**, 1493–1503 (2022).
7. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLOS Computational Biology* **16**, e1008120 (2020).
8. von Bartheld, C. S., Bahney, J. & Herculano-Houzel, S. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology* **524**, 3865–3895 (2016).
9. Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3**, (2009).
10. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol* **82**, 239–259 (1991).
11. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
12. 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).

13. Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat Genet* **56**, 605–614 (2024).

14. Haglund, A. *et al.* Cell state-dependent allelic effects and contextual Mendelian randomization analysis for human brain phenotypes. *Nat Genet* **57**, 358–368 (2025).

15. 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).