Title

Coordinated dysregulation of modular gene activity in human neuropathologies

ISBN

9798189297649

Author

Kang, Gugene

Publication Date

2026-09-03

Peer reviewed|Thesis/dissertation

Coordinated dysregulation of modular gene activity in human neuropathologies

by
Gugene Kang

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Developmental and Stem Cell Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Alex Pollen
2102D531BF2D416...

Alex Pollen
Chair

DocuSigned by:
Michael Oldham

Michael Oldham

DocuSigned by:
Todd Nystul

Todd Nystul

DocuSigned by:
Annette Molinaro
1B4CE53175EB453...

Annette Molinaro

Committee Members

Acknowledgments

I'd like to acknowledge my family, who gave unconditional support; Mike, who showed more patience than I probably deserved; and to all those who helped along the way.

Contributions

All work was conducted under the supervision of Michael C. Oldham. Chapter 1 is an original introduction written for this dissertation. Chapter 2 has been prepared as a manuscript (Kang & Oldham) for publication.

Citations:

Gugene Kang & Michael C. Oldham. Coordinated dysregulation of modular gene activity in human neuropathologies. (in preparation).

Coordinated dysregulation of modular gene activity in human neuropathologies

Gugene Kang

Abstract

Understanding which genes are reproducibly dysregulated in which cell types is foundational knowledge for efforts to slow or reverse pathologies. For neuropathologies, such efforts rely primarily on differential expression analysis of single-nucleus RNA-seq (snRNA-seq) data. However, this strategy suffers from experimental and statistical challenges that limit marker gene reproducibility. We describe a novel strategy called Covariation Projection Analysis (**CoPA**) that combines the power of bulk sampling with the precision of single-cell methods. By 'projecting' bulk gene coexpression modules onto pseudobulked snRNA-seq cell types, CoPA reveals the cellular origins of highly reproducible genomic programs and their relative importance among cell types. By comparing CoPA projection patterns between normal and pathological human brain samples using **d**ifferential CoPA (**dCoPA**), we identify gene coexpression modules that are uniformly and reproducibly dysregulated in specific neocortical cell types in Alzheimer's disease or schizophrenia. We share our findings through a novel web application called CoPA Cabana (https://oldhamlab.shinyapps.io/copacabana/).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

Development of single-cell and single-nucleus transcriptomics

The viability of extracting genomic information at the single-cell level was first demonstrated more than 30 years ago. PCR-based cDNA amplification techniques, including exponential[1] amplification, multiple-round linear amplification,[2] or a combination of the two,[3] enabled the amplification of cDNA from sub-picogram quantities of RNA.[4] These techniques were motivated by biological contexts that necessitated the study of single or small amounts of cells over population measurements. For example, inferring signaling network properties from population-level data can mask temporal fluctuations in intracellular signaling.[5] More practically, cell supply is inherently limited in certain contexts, e.g., early embryo development or stem cell biology, where small, distinct subsets of cells can serve pivotal biological functions. For example, Kurimoto et al. successfully analyzed single-cell gene expression via microarray from manually dissected E3.5 mouse blastocysts; Esumi et al. did the same with GABAergic neurons.[3,6]

Tang et al. applied next-generation sequencing[7] to cDNA amplified from a single mouse blastomere, effectively demonstrating the first example of single-cell RNAseq and achieving a 75% increase in the number of genes detected over microarray studies of multiple blastomeres.[8,9] Subsequently, several techniques were developed that improved throughput, sensitivity, and transcript coverage. Several protocols, including STRT-seq and CEL-seq, introduced different versions of single-cell barcoding: individual cells were sorted manually into each well of a 96-well plate and well-specific barcodes

were incorporated during reverse transcription, allowing for the pooling of barcoded cells for sequencing and subsequent computational demultiplexing to distinguish cell-specific transcripts.[10,11] Unique molecular identifiers (UMIs) were soon formalized as a universal method to enable pooling strategies that multiplicatively increase the throughput of amplification-based assays.[12] SMART-seq addressed the 5' bias of previously published protocols via a template-switching oligo incorporated during reverse transcription that allowed for full-length transcript reads.[13]

These plate-based methods were quickly superseded by droplet-based methods that increased throughput by orders of magnitude. Unlike previous methods, where cells were manually separated via pipette or FACS-sorted into individual wells of a 96- or 384-well plate, Macosko et al. devised a strategy to encapsulate individual cells alongside beads coated with barcode- and UMI-labeled oligonucleotide primers in aqueous droplets at nanoliter scale using a microfluidic device, allowing for the profiling of tens of thousands of cells in a single experiment[14]; Klein et al. created a similar approach called inDrop that used hydrogel beads to achieve greater capture efficiency[15]; both studies were published simultaneously in the same journal. Droplet-based techniques were eventually commercialized by 10x Genomics via their Chromium platform and became the dominant paradigm for high throughput single-cell sequencing.[16]

Technical difficulties associated with the availability and processing of certain tissues (e.g., brain, skeletal muscle, and adipose tissue) led to the preferential usage of single-nucleus sequencing over single-cell sequencing for those tissue contexts.[17,18] Single-nucleus sequencing, where isolated nuclei are sequenced rather than whole cell

lysates, possesses several technical advantages over single-cell sequencing: the quality of cDNA obtained from frozen tissue using whole cells is lower than when using nuclei,[19] and human brain tissue is widely available in frozen tissue archives while being comparatively difficult to obtain from fresh tissue biopsies. Moreover, human brain tissue is heterogeneous and susceptible to cell sampling bias: the ratio of excitatory to inhibitory neurons found by single-nucleus sequencing of the adult human cortex was significantly more accurate than single-cell sequencing of the same tissue type.[20-22] Finally, single-nucleus processing avoids cell isolation procedures that have been shown to increase expression of immediate early genes, mimicking the expression profile of activated neurons and potentially confounding downstream gene expression analyses.[23] However, these advantages come at the cost of the inability to sequence cytosolic RNA and lower RNA output overall.

The rapid and drastic reduction in cost-per-sequenced-cell has led to the creation of various cell atlas initiatives such as the Human Cell Atlas.[24] These efforts, spearheaded by international consortia with datasets contributed by labs worldwide, seek to provide comprehensive pan-tissue reference maps of single-cell identity, expression, and location across a range of species and disease conditions. Efforts specific to the human brain use single-nucleus sequencing as the default modality and are headlined by the Allen Brain Cell Atlas (https://brain-map.org/bkp/explore/abc-atlas), one public-facing output of the Brain Initiative Cell Census Network (BICCN) consortium (https://biccn.org). The human brain transcriptomic data represented by this resource consists of more than three million nuclei spanning 100 anatomically distinct regions across three neurotypical adult human brains,[25] with additional nuclei included from

various cortical regions of two additional brains.[26] These regions include the dorsolateral prefrontal cortex, anterior cingulate cortex, middle temporal gyrus, angular gyrus, and primary motor, somatosensory, auditory, and visual cortices. The majority of these neurotypical adult cortical nuclei were sequenced using v3 of the 10x Chromium droplet-based platform, while a subset was sequenced at higher depth using v4 of the SMART-seq platform. Cortical tissue samples were also spatially resolved using MERFISH, an *in situ* hybridization-based single-cell imaging technique.[27] Prior to sequencing, nuclei were sorted according to NeuN signal and were combined at a fixed ratio of 80:20 NeuN+ to NeuN- nuclei, effectively enforcing an 80% neuron to 20% non-neuron ratio per sample.

To classify cell types, the Allen Brain Cell Atlas constructed a reference cellular taxonomy using studies of the mammalian primary motor cortex.[28] This mammalian motor cortex taxonomy was derived from seven mouse single-cell/single-nucleus RNAseq datasets, a single-nucleus methylcytosine sequencing (snmC-seq2) dataset, and a single-nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq) dataset,[29] alongside single-nucleus RNAseq data from human and marmoset primary motor cortex.[30] First, mouse transcriptomic and epigenomic data were integrated using two computational methods, one based on non-negative matrix factorization[31] and one utilizing a mutual nearest neighbor approach,[32] to create a consensus molecular mouse taxonomy consisting of 116 total clusters. This mouse taxonomy was then integrated with a similar taxonomy produced from human and marmoset primary motor cortex using Seurat's SCTransform pipeline[33] to produce the final cross-species motor cortex taxonomy of 45 cell-types, (called "t-types"). T-types

were broadly classified into 24 GABAergic, 13 glutamatergic, and 8 non-neuronal types. GABAergic types were further subdivided based on developmental origin: caudal ganglionic eminence (CGE)-derived (*Lamp5, Vip, Sncg*) or medial ganglionic eminence (MGE)-derived (*Pvalb, Sst, Sst Chodl*). Glutamatergic neurons were subdivided based on layer (2 to 6) and projection type (intratelencephalic, extratelencephalic, corticothalamic, and near-projecting). Nuclei from the neurotypical adult human cohort were assigned to 24 of the 45 t-types based on transcriptomic similarity using Seurat's label transfer.[26] These 24 cell types, or "subclasses", were further clustered into more granular cell types ("supertypes"), ranging from 120 to 142 supertypes depending on the cortical region.

The Allen Brain Cell Atlas also includes data from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD), a cohort derived from 84 donors exhibiting varying degrees of Alzheimer's disease (AD) pathology. Over three million nuclei from the middle temporal gyrus and the dorsolateral prefrontal cortex were collected and used for single-nucleus RNAseq using the 10x Chromium V3 platform, sn-ATACseq, and MERFISH profiling.[34] Similarly to other atlas studies, nuclei were sorted and combined at a 70:30 NeuN+ to NeuN- ratio. Cell type labels (subclass/supertype) were transferred using labels from the neurotypical adult human cohort.[26] Study authors assigned each donor a pseudoprogression score indicating the extent of AD progression. The score utilized a Bayesian model that inferred disease progression from quantitative neuropathological measurements, including pTau, amyloid beta, NeuN, pTDP-42, alpha synuclein, IBA1, and GFAP expression. Each donor was assigned a continuous score ranging from 0 to 1, ostensibly offering greater granularity than traditional AD staging

metrics such as Braak stages, Thal phases, CERAD score, or ADNC score. Using this pseudoprogression score, the authors utilized scCODA,[35] a Bayesian model for cell type composition analysis, to estimate the relative change in abundance of cell types across AD progression. The authors included sex, race, 10x chemistry, and APOE4 status as covariates in the model. Based on the modeling results, the authors proposed a two-epoch model for AD progression. The first epoch involves selective SST neuron and oligodendrocyte loss, upregulation of disease-associated microglia and protoplasmic astrocytes, and compensatory upregulation of oligodendrocyte precursor cells (OPCs). The late epoch, characterized by an exponential increase in pTau and amyloid beta plaques alongside the emergence of cognitive impairment, involves broad neuronal loss among L2/3 intratelencephalic (L2/3 IT) neurons, PVALB interneurons, VIP interneurons, OPCs, with a continued increase in the relative abundance of astrocytes and microglia. The authors reported that 26% of supertypes showed similar changes in relative abundance when ADNC or cognitive status were used as the explanatory variable in the modeling instead of the pseudoprogression score. The authors assessed reproducibility in two reprocessed external studies[36,37] and observed similar changes in cell type abundance for some cell types (SST, L2/3 IT, microglia) but not others (oligodendrocytes). The authors argued that these discrepancies were driven by fewer late-stage AD donors and fewer nuclei per donor.

Limitations of single-cell and single-nucleus transcriptomics

The technical complexity of single-cell and single-nucleus data production and analysis revealed several challenges, including tissue dissociation artifacts, cell

sampling bias, ambient RNA contamination, doublets, dropout events, and pseudoreplication bias.

Tissue dissociation artifacts and cell sampling bias

Unlike bulk tissue sequencing, where genomic material is extracted from homogenized tissue samples, single-cell sequencing requires an initial tissue dissociation step to create a single-cell suspension via enzymatic and/or mechanical disruption. The activation of stress-related genes by tissue dissociation is well described across a variety of different tissue contexts.[23,38] For example, Denisenko et al. analyzed a variety of tissue dissociation and storage protocols using adult mouse kidney tissue and observed systematic gene expression artifacts and sampling bias regardless of method.[39] Dissociating tissue at 37°C resulted in activation of immediate early response genes (*Fosb*, *Fos*, *Jun*, *Junb*, *Atf3*) as well as heat shock genes (*Hspa1a, Hspa1b*). Aberrant expression of these stress response genes were largely attenuated upon using cold tissue dissociation, but the authors found significant differences in cell type abundance between warm and cold dissociation protocols for eight cell types including podocytes and endothelial cells, suggesting cell-type-specific survival rates for a given protocol.

Single-nucleus sequencing involves extraction of nuclei from lysed tissue, avoiding tissue dissociation and expression of stress response genes.[23,40] However, single-nucleus sequencing also exhibits cell-sampling bias: Denisenko et al. reported a significant underrepresentation of immune cells in single-nucleus sequencing relative to single-cell[39]; Machado et al. reported a significant decrease of endothelial cells in mouse skeletal muscle and liver in single-nucleus sequencing compared to single-cell.[40]

Other methods to circumvent tissue dissociation-related artifacts have emerged, but come with tradeoffs.[41] *In situ* fixation preserves RNA integrity and composition, but is currently not compatible with 10x Chromium platforms. Selective inhibition of transcriptional machinery can reduce procedure-related spurious gene expression, but the homogeneous diffusion of inhibitors can be difficult to guarantee, and inhibitors do not account for post-transcriptional modifications. Conditional labeling of cell-type-specific RNA prior to dissociation offers a precision approach to avoid dissociation artifacts but is limited by the practical application of transgenics to diverse experimental settings.

Ambient RNA contamination

Droplet-based platforms rely on the assumption that each droplet contains RNA from a single cell or nucleus, but in practice droplets incorporate free-floating ambient RNA from lysed cells that can distort downstream analyses of gene expression. In response various *post hoc* algorithms have been developed to remove ambient RNA signatures *in silico.*[42,43] Some methods (CellBender,[44] SoupX,[45] FastCAR,[46] CellClear,[47] scCDC[48]) infer ambient RNA expression from empty droplets, while others (DecontX,[49] scAR[50]) model ambient RNA expression from the filtered data alone. Benchmarking results using various real and synthetic datasets indicate that no single method outperforms all others across all datasets and evaluation criteria.[43] Choice of decontamination algorithm therefore presents another source of preprocessing variability that can impact downstream results.

Doublets

In the context of Drop-seq-based sequencing protocols, doublets (or more generally multiplets) are droplets containing genomic material from two or more cells/nuclei. Doublets can produce spurious results in downstream analyses and necessitate the usage of various computational and experimental methods for doublet removal. Doublet rate typically depends on cell loading density: 10x reports a scaling rate of approximately ~0.8% per 1000 cells loaded on their Chromium v3 platform.[51] This encourages conservative cell loading thresholds, limiting sequencing throughput.

Commonly used doublet identification algorithms include scDblFinder,[52] DoubletFinder,[53] and Scrublet.[54] These and other methods operate by generating artificial doublets by combining expression profiles from two cells, then training a classifier to distinguish real droplets from artificial doublets.[55] These approaches often require user input for certain parameters such as expected doublet rate, which can be difficult to estimate in practice. Moreover, these approaches are sensitive to heterotypic doublets (i.e., doublets that consist of two transcriptionally divergent cell types) but fail to accurately detect homotypic doublets (i.e. doublets that consist of two cells of the same type). Several experimental approaches have been developed to resolve doublets,[51,56,57] but these methods are costly and cannot be used to identify doublets from existing data.[55]

Dropouts

Single-cell data are notoriously sparse due to various biological and technological factors.[58] While the presence of zeroes may reflect a true lack of expression of a gene for a given cell, it can also reflect the stochastic and “bursty”

nature of transcription.[59] The expression of a gene depends on whether the gene is in an “active” transcriptional state as well as the rate of mRNA degradation.[60] Technical reasons for zeroes include the following: the stochasticity and highly variable efficiency of reverse transcription, which can be as low as 20%[61]; the sequencing depth of the experiment, where genes with low expression may fail to reach thresholds for efficient cDNA amplification; and various technical biases associated with cDNA amplification including GC bias[62] and inefficiency related to barcode, primer, and adapter design.[63] The situation is even worse for single-nucleus studies, since all cytoplasmic RNA is lost.

As a result of this sparsity, single-cell and single-nucleus RNA-seq data are less well suited to gene expression correlation analyses, which has inspired various efforts to algorithmically impute gene expression.[64] Experiments using synthetic data suggest that cell-level analyses such as clustering are more robust to zeroes than gene-level analyses such as differential expression.[58]

Pseudoreplication bias

In a single-cell experiment with multiple donors, cells from the same individual are not independent observations and cell-level statistical methods (e.g. differential expression analysis) that ignore this are prone to pseudoreplication bias and can underestimate variance as a result.[65] Creating pseudobulk samples by aggregating raw counts from cellular populations within each biological replicate and then performing downstream analysis addresses this issue, but this has the side effect of drastically diluting the statistical power of the single-cell data. Reanalysis of a single-nucleus study[66] using this pseudobulk approach before differential expression analysis

dramatically reduced the number of genes reported as significantly differentially expressed compared to what the study had originally reported.[67]

Leveraging bulk transcriptomic data to infer gene-level relationships

The sparsity and overdispersion of single-cell data complicate analytical efforts at the gene level. Analyses of synthetic single-cell datasets suggest that differential expression analyses are less robust to sparsity than clustering and other cell-level analyses.[58] Methods developed to analyze gene co-expression networks in bulk data perform poorly when applied to single-cell data.[68] Commonly used measures of gene association such as Pearson correlation, euclidean distance, and mutual information performed poorly in benchmark analyses.[69] Several methods have been developed to create gene co-expression networks from single-cell data that rely on a variety of modeling techniques and transformative adjustments to control for single-cell-specific technical considerations, but benchmark analyses have shown that these methods still suffer from type I error and estimation bias.[70]

In contrast, gene co-expression is well-studied in the context of bulk transcriptomics and has been shown to be robust, scalable, and specific despite a lack of cell-level resolution. Shortly after the widespread adoption of genome-wide expression profiling, several groups demonstrated that clustering genes according to the similarity of their expression patterns could recreate patterns of shared biological function. Eisen et al. analyzed gene expression data for *Saccharomyces cerevisiae* using DNA microarrays and demonstrated that hierarchically clustering genes using correlation as a distance metric resulted in genes with similar function (e.g. stress, mitochondrial, proteasome, histone) clustering together.[71] Butte and Kohane organized

genes into an explicit network with genes as nodes and edges representing similarity via mutual information.[72] Stuart et al. also utilized a graph-based gene network representation but expanded the analysis to a cross-species approach where edges between nodes were only included if they replicated across humans, flies, worms, and yeast.[73] Stuart et al. also introduced the concept of “modules”, defined as tightly clustered groups of genes that were represented in the co-expression graph as highly interconnected neighborhoods. The authors demonstrated via Gene Ontology enrichment that these modules represented functional units of the genome representing conserved biological processes.

These network concepts were formalized as an analytical framework named Weighted Gene Co-expression Network Analysis (WGCNA) in 2005.[74] Several analytical choices distinguish WGCNA from previous studies. Whereas Stuart et al. used a hard significance cutoff to assign edges to nodes in a binary fashion, WGCNA used a soft thresholding approach to weight gene similarities so that the overall co-expression network approximated a scale-free topology, where a small number of highly connected “hub” genes are surrounded by many nodes with fewer links; this reflects the observation that many biological networks are scale-free in nature.[75] From this weighted similarity matrix, WGCNA then calculates the topological overlap measure (TOM) for each pair of nodes, a higher-order similarity measure which incorporates information about shared neighbors into the weighted similarity measure, increasing robustness and reducing the impact of spurious single edges. To discover modules, this matrix of gene-gene TOM similarity values was hierarchically clustered, then algorithmically cut at various adaptive heights to produce lists of modules. WGCNA thus synthesized various

network concepts to create a statistically-motivated approach to detect gene co-expression modules from genome-wide transcriptomic data. The WGCNA workflow was consolidated into a single R package published in 2008.[76]

WGCNA's formal approach to module discovery allowed for a shared technical language to emerge to describe the various properties of modules. Each module consisted of a list of genes called "seed genes" that were summarized by calculating the first principal component across all seed gene expression vectors; this summary vector was referred to as the "module eigengene".[77] Summarizing modules in this fashion enabled a number of downstream analyses and applications. Modules could be merged if the similarity between their module eigengenes exceeded a certain threshold. All genes in the network could be ranked according to their similarity with a given module eigengene; this ranking is referred to as a gene's k_{ME} value. Top-ranked genes by k_{ME} were of particular biological interest and represented hub genes for a given module. Finally, module eigengenes themselves could be hierarchically clustered into a "meta-network" that represents a higher-order organization of the transcriptome.[77]

Subsequent studies applied WGCNA to various tissue contexts and demonstrated the power of a module-based organization of the transcriptome to discover highly conserved and specific signatures of biological processes and cell types. Horvath et al. analyzed two independent glioblastoma multiforme (GBM) microarray datasets spanning 120 samples and discovered that *ASPM* was a hub gene associated with a module that was also present in breast cancer that represented a "meta-signature" for undifferentiated cancer; *ASPM* inhibition suppressed tumor proliferation and neural stem cell proliferation.[78] Oldham et al. applied WGCNA to four

human brain microarray datasets spanning 160 human brain control samples representing the cortex, caudate nucleus and cerebellum.[79] The analysis yielded modules enriched with markers of major brain cell classes in each of the three regions analyzed. Other modules were found that were enriched with markers for additional cell types such as meningeal cells and Purkinje neurons, as well as modules related to ribosomal, mitochondrial, and synaptic function, demonstrating that gene co-expression is sensitive enough to detect cell-type-specific signatures despite lack of cell-type-specific resolution. In addition, gene co-expression analysis distinguishes itself from differential expression by providing a framework to describe the inherent hierarchical architecture of the transcriptome that encompasses both cell-type-specific signatures and non-specific biological processes.

Kelley et al. expanded the work of Oldham et al. and applied co-expression analysis at scale, analyzing 7221 samples representing 840 neurotypical human individuals across 19 broad neuroanatomical regions sourced from microarray and bulk RNAseq datasets.[80] These data were used to generate modules representing highly conserved transcriptomic signatures that were both sensitive and specific for four major brain cell classes: astrocytes, oligodendrocytes, microglia, and neurons (AOMN). Kelley et al. demonstrated that many of the genes highly associated with these modules were not described in the literature, and showed an example of a top microglial gene (*APBB1IP*) that co-stained nearly perfectly with a canonical microglial marker (*AIF1*) despite having no associated literature at the time. These AOMN-associated genes overlapped partially but not fully with cell-class markers obtained from adult human brain snRNAseq datasets,[81,82] and moreover the genes that were implicated as markers

in the bulk context but not in the single-nucleus context had significantly lower expression in the single-nucleus data than genes that were concordant markers across both data modalities. Kelley et al. highlighted the inherent suitability of bulk expression data over single-cell data for gene co-expression analyses and used cell-type-specific modules to discover novel cell-type-specific differences in various biological and disease contexts.

References

1. Brady, G. Representative in Vitro cDNA Amplification From Individual Hemopoietic Cells and Colonies. (1990). https://www.academia.edu/4577596

2. Van Gelder, R.N. et al. Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A 87, 1663-1667 (1990).

3. Kurimoto, K. et al. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. Nucleic Acids Res 34, e42 (2006).

4. Iscove, N.N. et al. Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. Nat Biotechnol 20, 940-943 (2002).

5. Korobkova, E. et al. From molecular noise to behavioural variability in a single bacterium. Nature 428, 574-578 (2004).

6. Esumi, S. et al. Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors. Neurosci Res 60, 439-451 (2008).

7. Mardis, E.R. The impact of next-generation sequencing technology on genetics. Trends Genet 24, 133-141 (2008).

8. Maekawa, M. et al. Requirement for ERK MAP kinase in mouse preimplantation development. Development 134, 2751-2759 (2007).

9. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6, 377-382 (2009).

10. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Rep 2, 666-673 (2012).

11. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 21, 1160-1167 (2011).

12. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 9, 72-74 (2012).

13. Ramsköld, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol 30, 777-782 (2012).

14. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161, 1202-1214 (2015).

15. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187-1201 (2015).

16. Zheng, G.X.Y. et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 8, 14049 (2017).

17. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol 38, 737-746 (2020).

18. Grindberg, R.V. et al. RNA-sequencing from single nuclei. Proc Natl Acad Sci U S A 110, 19802-19807 (2013).

19. Krishnaswami, S.R. et al. Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. Nat Protoc 11, 499-524 (2016).

20. Bakken, T.E. et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One 13, e0209648 (2018).

21. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A 112, 7285-7290 (2015).

22. Lake, B.B. et al. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 352, 1586-1590 (2016).

23. Lacar, B. et al. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. Nat Commun 7, 11022 (2016).

24. Regev, A. et al. The Human Cell Atlas. Elife 6, e27041 (2017).

25. Siletti, K. et al. Transcriptomic diversity of cell types across the adult human brain. Science 382, eadd7046 (2023).

26. Jorstad, N.L. et al. Transcriptomic cytoarchitecture reveals principles of human neocortex organization. Science 382, eadf6812 (2023).

27. Chen, K.H. et al. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015).

28. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature 598, 86-102 (2021).

29. Yao, Z. et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature 598, 103-110 (2021).

30. Bakken, T.E. et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature 598, 111-119 (2021).

31. Welch, J.D. et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. Cell 177, 1873-1887.e17 (2019).

32. Luo, C. et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. Cell Genom 2, 100107 (2022).

33. Stuart, T. et al. Comprehensive integration of single-cell data. Cell 177, 1888-1902.e21 (2019).

34. Gabitto, M.I. et al. Integrated multimodal cell atlas of Alzheimer's disease. Nat Neurosci 27, 2366-2383 (2024).

35. Büttner, M. et al. scCODA is a Bayesian model for compositional single-cell data analysis. Nat Commun 12, 6876 (2021).

36. Green, G.S. et al. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. Nature 633, 634-645 (2024).

37. 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).

38. Machado, L. et al. In Situ Fixation Redefines Quiescence and Early Activation of Skeletal Muscle Stem Cells. Cell Rep 21, 1982-1993 (2017).

39. Denisenko, E. et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol 21, 130 (2020).

40. Machado, L. et al. Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation. Cell Stem Cell 28, 1125-1135.e7 (2021).

41. Machado, L., Relaix, F. & Mourikis, P. Stress relief: Emerging methods to mitigate dissociation-induced artefacts. Trends Cell Biol 31, 888-897 (2021).

42. Caglayan, E., Liu, Y. & Konopka, G. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. Neuron 110, 4043-4056.e5 (2022).

43. Cargnelli, C.B., Nielsen, J.V. & Madsen, J.G.S. Benchmarking computational decontamination of ambient RNA. bioRxiv 2026.01.13.699237 (2026).

44. Fleming, S.J. et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. Nat Methods 20, 1323-1335 (2023).

45. Young, M.D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience 9 (2020).

46. Berg, M. et al. FastCAR: fast correction for ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. BMC Genomics 24, 722 (2023).

47. Huang, W., Hu, L., Qian, Z. & Fan, J. CellClear: Enhancing Single-cell RNA Data Quality via Biologically-Informed Ambient RNA Correction. bioRxiv 2024.08.05.606571 (2024).

48. Wang, W. et al. scCDC: a computational method for gene-specific contamination detection and correction in single-cell and single-nucleus RNA-seq data. Genome Biol 25, 136 (2024).

49. Yang, S. et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol 21, 57 (2020).

50. Sheng, C. et al. Probabilistic modeling of ambient noise in single-cell omics data. bioRxiv 2022.01.14.476312 (2022).

51. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol 19, 224 (2018).

52. Germain, P.L. et al. Doublet identification in single-cell sequencing data using scDblFinder. F1000Res 10, 979 (2022).

53. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst 8, 329-337.e4 (2019).

54. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281-291.e9 (2019).

55. Xi, N.M. & Li, J.J. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. Cell Syst 12, 176-194.e6 (2021).

56. Kang, H.M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 36, 89-94 (2018).

57. McGinnis, C.S. et al. MULTI-seq: Sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods 16, 619-626 (2019).

58. Jiang, R., Sun, T., Song, D. & Li, J.J. Statistics or biology: The zero-inflation controversy about scRNA-seq data. Genome Biol 23, 31 (2022).

59. Raj, A. et al. Stochastic mRNA Synthesis in Mammalian Cells. PLoS Biol 4, e309 (2006).

60. Paszek, P. Modeling stochasticity in gene regulation: Characterization in the terms of the underlying distribution function. Bull Math Biol 69, 1567-1601 (2007).

61. Schwaber, J., Andersen, S. & Nielsen, L. Shedding light: The importance of reverse transcription efficiency standards in data interpretation. Biomol Detect Quantif 17, 100077 (2019).

62. Silverman, J.D., Roche, K., Mukherjee, S. & David, L.A. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J 18, 2789-2798 (2020).

63. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A 109, 1347-1352 (2012).

64. van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell 174, 716-729.e27 (2018).

65. Squair, J.W. et al. Confronting false discoveries in single-cell differential expression. Nat Commun 12, 5692 (2021).

66. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570, 332-337 (2019).

67. Murphy, A.E., Fancy, N. & Skene, N. Avoiding false discoveries in single-cell RNA-seq by revisiting the first Alzheimer's disease dataset. Elife 12, RP90214 (2023).

68. Chen, S. & Mar, J.C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics 19, 232 (2018).

69. Skinnider, M.A., Squair, J.W. & Foster, L.J. Evaluating measures of association for single-cell transcriptomics. Nat Methods 16, 381-386 (2019).

70. Su, C. et al. Cell-type-specific co-expression inference from single cell RNA-sequencing data. Nat Commun 14, 4846 (2023).

71. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95, 14863-14868 (1998).

72. Butte, A.J. & Kohane, I.S. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 418-429 (2000).

73. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science 302, 249-255 (2003).

74. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4, Article17 (2005).

75. Ravasz, E. et al. Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551-1555 (2002).

76. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).

77. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. BMC Syst Biol 1, 54 (2007).

78. Horvath, S. et al. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci U S A 103, 17402-17407 (2006).

79. Oldham, M.C. et al. Functional organization of the transcriptome in human brain. Nat Neurosci 11, 1271-1282 (2008).

80. Kelley, K.W., Nakao-Inoue, H., Molofsky, A.V. & Oldham, M.C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nat Neurosci 21, 1171-1184 (2018).

81. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 14, 955-958 (2017).

82. Lake, B.B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol 36, 70-80 (2018).

Chapter 2: Coordinated dysregulation of modular gene activity in human neuropathologies

Introduction

Understanding how gene activity differs among neurobiological cell types in health and disease is a major goal of neuroscientific research. Today, most studies of normal and pathological human brain samples address this problem through differential expression (DE) analysis of cell types identified through single-nucleus RNA-seq (snRNA-seq). Although conceptually straightforward, snRNA-seq requires many experimental steps and analysis choices, each of which can introduce bias, inflate type I and II errors, and impair the reproducibility of marker genes. Well-known sources of experimental bias and error include loss of cytosolic mRNA, buffering of stochastic transcription by the nucleus,[1,2] variable capture efficiency of nuclei and transcripts,[3-5] multiplets,[6-9] contamination with ambient RNA,[10-13] and artifacts from cDNA amplification.[14] Importantly, these factors are much more likely to distort marker gene identification than cell type identification, since the former relies on a single noisy gene expression vector while the latter relies on thousands. Furthermore, investigator choices can dramatically alter the results of DE analysis. For example, the choice of DE algorithm – and even the software package version – can produce very different lists of marker genes from the same underlying data.[15] Combined with concerns about pseudoreplication bias[16-18] and statistical power – particularly for low-expressed genes that dominate most snRNA-seq datasets[19,20] – it is important to evaluate the

reproducibility of marker genes identified by snRNA-seq and explore alternative strategies for studying pathological gene activity.

Genes do not function in isolation, but rather as members of coordinated genomic programs that support distinct cellular functions. We and others have shown that genome-wide coexpression analysis of bulk tissue samples is a powerful approach for revealing highly reproducible programs (or 'modules') of genomic activity,[21-28] since variation in the cellular composition of bulk tissue samples inevitably drives covariation of optimal markers for cell types and states at scale. Because one bulk sample may represent several million cells, large bulk datasets often represent *billions* of cells, providing enormous statistical power to survey the population structure of genomic activity while avoiding most sources of experimental bias and error that impact snRNA-seq studies. On the other hand, the cellular origins of most bulk coexpression modules are unknown. The complementary strengths and weaknesses of bulk and snRNA-seq data suggest their integration might clarify differences in genomic activity among neurobiological cell types in health and disease.

Here we introduce a novel approach called **Co**variation **P**rojection **A**nalysis (**CoPA**) that combines the power of bulk coexpression with the precision of single-cell analysis to reveal the cellular origins of highly reproducible modules of genomic activity. We focus on adult human neocortex and its 24 major cell types (or subclasses), which have been identified and validated as distinct transcriptional clusters in multiple snRNA-seq studies.[29-32] Using CoPA, we show that the overwhelming majority of gene coexpression modules in adult human neocortex are not cell-type-specific, but that subtle differences in the expression levels of coexpressed genes among cell types are

nevertheless extremely reproducible. By comparing module projection patterns between snRNA-seq datasets derived from normal and pathological human brain samples using **d**ifferential CoPA (**dCoPA**), we identify gene coexpression modules that are uniformly and reproducibly dysregulated in specific neocortical cell types from patients with Alzheimer's disease or schizophrenia. We share our results through a novel web application called CoPA Cabana (https://oldhamlab.shinyapps.io/copacabana/), which also enables CoPA for genome-wide coexpression networks in OMICON (https://theomicon.ucsf.edu), which are derived from standardized analysis of 116 datasets representing >17K normal and neoplastic human brain samples.

Results

Human neocortical subclass markers vary across snRNA-seq datasets and DE methods

We first examined the reproducibility of human neocortical subclass markers identified by DE analysis of snRNA-seq datasets. We focused on two large studies[29,30] that surveyed 253,177 nuclei from the dorsal frontal cortex (DFC) and 273,971 nuclei from the middle temporal gyrus (MTG) of 12 neurotypical adult donors using 10x Chromium v3 (Cv3). Both studies annotated cell types using the Brain Initiative Cell Census Network (BICCN) reference taxonomy, which defines 24 subclasses and ~130-150 'supertypes'.[32] Because subclasses are more clearly defined than supertypes and comprise the neocortex in more reproducible proportions (**Fig. S2.1**), we focused most analysis at this level. To avoid pseudoreplication bias and mitigate false discoveries,[16-18] we summed expression levels for each gene over all subclass nuclei from each donor prior to donor-level DE analysis with edgeR[33] glmLRT or DESeq2[34] (test = 'LRT'), which were among the top performing methods in a benchmark study of DE analysis for

single-cell datasets.[18] We compared each subclass to all other subclasses and defined marker genes as those that were significantly upregulated (FDR <.05) and unique (i.e., significant for only one subclass).

We evaluated the reproducibility of marker genes identified by edgeR or DESeq2 in different datasets and brain regions (**Fig. 2.1, Fig. S2.2**). Over all subclasses and both DFC datasets, on average only 33.7% of markers were identified by both methods (**Fig. 2.1a,b**). In total, we identified only 110 genes as reproducible subclass markers in both datasets using both methods (**Fig. 2.1c**), and 79% of these (n=87) were non-neuronal (**Fig. 2.1d**). No neuronal subclass had more than five reproducible markers, and nine neuronal subclasses had none (**Fig. 2.1d**). To explore the basis for these findings, we formed subclass metacells from each dataset by averaging gene expression for all subclass nuclei, then clustered metacells by genome-wide dissimilarity (1-cor). This analysis revealed that most neuronal subclasses were extremely similar, with average genome-wide correlations of 0.94–0.95 among excitatory subclasses, 0.91–0.92 among inhibitory subclasses, and 0.9–0.91 among all neuronal subclasses; in contrast, non-neuronal subclasses were much more heterogeneous (**Fig. 2.1e,f**). Similar results were observed in MTG (**Fig. S2.2**). These findings underscore challenges with DE analysis of snRNA-seq data, consistent with previous reports.[35-37]

Genome-wide expression levels are poorly explained by cell-type abundance in pseudobulk datasets

To further analyze the contributions of individual cell types to the expression levels of individual genes, we created pseudobulk samples using snRNA-seq data from

individual donors and cortical regions.[30] To create each pseudobulk sample, we randomly sampled nuclei from all subclasses, recorded their identities, and summed the expression levels for each gene over all sampled nuclei. Using large pseudobulk datasets (n=1,518 samples), we then performed multiple linear regression to model gene expression as a function of subclass abundance, with each gene serving as the response variable, subclass abundance vectors (n=24) as predictors, and adjusted R^2 and root mean-squared error (RMSE) as model outputs (**Fig. 2.2a**). We found that variation in subclass abundance explained only 7.6–8.2% of genome-wide expression variation (n=17,753 genes) in DFC samples from three donors (**Fig. 2.2b**). To determine if more granular cell type definitions would produce better results, we constructed pseudobulk datasets from the same snRNA-seq data by randomly sampling supertypes (n=136) instead of subclasses and repeated the analysis. We found that variation in supertype abundance explained only 8.8–9.6% of genome-wide expression variation (**Fig. 2.2b**). We repeated these analyses for additional cortical regions from the same donors. In MTG, we found that 7.5–8.1% of genome-wide expression variation was explained by variation in subclass abundance and 8.5–9.5% by supertype abundance (**Fig. S2.3a**). In primary visual cortex (V1), these numbers were even lower (5.1–6.6% for subclass and 6.7–8.2% for supertype; **Fig. S2.3c**).

To validate these findings in a second snRNA-seq dataset,[29] we repeated pseudobulk modeling using DFC samples from nine additional donors and found that 6–10% of genome-wide expression variation was explained by variation in subclass abundance and 7.3–13% by supertype abundance (**Fig. S2.4**). To determine whether deeper sequencing of nuclei would produce better results, we analyzed Smart-SEQ v4

(SSv4) snRNA-seq data[30] from the same individuals and regions shown in **Fig. 2.2b** (DFC) and **Fig. S2.3a** (MTG). Despite >100x the read depth of Cv3 snRNA-seq data, only 4.9–6.6% of genome-wide expression variation in DFC or MTG samples analyzed with SSv4 was explained by variation in subclass abundance and only 6.9%–13% by supertype (**Fig. S2.5**).

Previous studies have demonstrated that very low gene expression measurements in single cells (< 1 UMI / cell, on average) are generally unreliable for inferring meaningful cell-to-cell variation in expression levels.[19,20] In every pseudobulk analysis, we observed that modeling performance was significantly worse for genes expressed below this threshold, which represented 74.2–82.5% of all genes in Cv3 datasets and 18.8–28.9% of all genes in SSv4 datasets (**Figs. 2.2c, S2.3b&d, S2.4b, S2.5b&d)**. We further explored this finding at the level of individual subclass markers and observed better modeling performance for more highly expressed marker genes (**Fig. 2.2d**). However, even for the most deeply sequenced genes and datasets (SSv4), modeling performance was still on average quite poor. Collectively, these results indicate that human neocortical cellular identities are poor predictors for the expression levels of most genes in current snRNA-seq datasets.

Covariation Projection Analysis (CoPA) reveals highly conserved modular gene activity in bulk and snRNA-seq datasets

The difficulty identifying reproducible marker genes for human neocortical subclasses may reflect the noisy nature of individual gene expression vectors in snRNA-seq data. However, genome-wide similarities among subclass metacells (**Fig. 2.1e-f, Fig. S2.2e-f**) strongly suggest that most gene activity in human neocortex is not

cell-type-specific, which begs two important questions: How do shared genomic programs vary among neocortical cell types? And how are shared genomic programs impacted by disease? To address these questions, we searched for highly conserved gene coexpression modules in a large bulk RNA-seq dataset comprised of samples from six studies[38-44] that surveyed genomic activity in neurotypical adult human DFC. After preprocessing and batch correction,[45,46] the combined dataset consisted of 1,518 samples representing billions of cells, providing enormous statistical power to resolve highly reproducible gene coexpression modules in human neocortex.

We used an iterative version of a previously described module detection algorithm[22] (**Fig. 2.3a**) to identify 1,016 fine-grained gene coexpression modules, which were numbered in the order they were identified. Each module was summarized by its first principal component, or module eigengene (ME), and the similarity between each gene and each ME, or k_{ME}, was quantified by Pearson correlation.[47] Following these steps, we produced a 'snapshot' for each module featuring several visualizations (**Fig. 2.3b-f**). First, we examined module coherence by calculating the percent variance explained by the ME and plotting z-scored expression patterns of the top 10 module genes (ranked by k_{ME}) for 100 randomly selected samples from our combined dataset (**Fig. 2.3b**). Second, we assessed module reproducibility by plotting the distributions of pairwise correlations for module seed genes among samples from each of the input datasets we analyzed. These distributions were compared to empirical null distributions of pairwise correlations for randomly sampled genes (equivalent in number to module size) (**Fig. 2.3c)**. Third, we characterized module functions via enrichment analysis with

a large collection of gene sets from the Molecular Signatures Database[48] and OMICON (**Fig. 2.3d**).

To elucidate the cellular origins of bulk coexpression modules, we projected module identities onto four snRNA-seq datasets from normal adult human DFC.[29-31,49] These datasets collectively represent 1,317,955 nuclei from 74 individuals sequenced to an average depth of 11,778 UMIs per nucleus. All datasets but one used the BICCN reference taxonomy[32] for cell type annotation. To harmonize annotations for the remaining dataset,[49] we assigned each nucleus to a BICCN subclass based on its maximum correlation with subclass metacells from Gabitto et al.[29] (**Fig. S2.6**). To project bulk coexpression modules onto snRNA-seq subclasses, we first calculated the mean expression level (log-transformed UMI counts) for all module genes in all nuclei belonging to each subclass. This analysis revealed a common pattern for many modules, with highest expression in excitatory neurons, intermediate expression in inhibitory neurons, and lowest expression in non-neurons (**Fig. 2.3e**). Averaging genome-wide expression levels for each subclass confirmed this overall pattern (**Fig. S2.7**), which was consistent across cortical regions and reflects differences in the sizes of subclass nuclei and their overall transcriptional outputs. We therefore normalized each module projection by dividing the mean expression of module genes by the mean expression of all genes for each subclass and scaling all subclass values to lie between (0,1). This 'relative expression index' (REI) therefore quantifies the mean expression of module genes relative to the mean expression of all genes for each subclass (**Fig. 2.3f**).

Snapshots of four 'archetypal' modules are shown in **Fig. 2.4a-p**, representing gene expression in all subclasses (column 1), all neuronal subclasses (column 2), a

single subclass (column 3), or multiple non-neuronal subclasses (column 4). Although module coherence declined as expected with increasing module number (**Fig. 2.4a-d**), module reproducibility generally remained quite high (**Fig. 2.4e-h**). Many modules were significantly enriched with multiple gene sets (**Fig. 2.4i-l**) whose functions were consistent with module gene identities and projection patterns. We also observed that module projection patterns were very consistent for independent snRNA-seq datasets; in many cases, even subtle differences in REI values were extremely reproducible (**Fig. 2.4m-p**). Overall, 12,908 genes (68% of our input dataset) were identified as module seed genes (mean: ~13 seed genes / module) (**Fig. 2.4q**). Using an expanded module definition based on k_{ME} values ('topmodposbc'), 17,972 genes (95% of our input dataset) were assigned to modules (mean: ~18 genes / module) (**Fig. 2.4q**). The mean pairwise correlation of module seed genes ranged from ~0.95 for the first module to ~0.37 for the last module (**Fig. 2.4r**). Over all modules, the mean correlation of REI projection patterns among all snRNA-seq datasets was 0.87 (**Fig. 2.4s**). All module snapshots can be browsed or searched on the CoPA Cabana web site that accompanies this study (https://oldhamlab.shinyapps.io/copacabana/).

To systematically categorize module projection patterns and visualize the overall organization of gene activity in human neocortical subclasses, we performed consensus clustering of CoPA projection patterns for DFC snRNA-seq datasets from Jorstad et al.[30] and Gabitto et al.[29] (**Fig. 2.5a**). Specifically, by clustering modules using 1 – the minimum Pearson correlation of REI projections from both datasets, we identified 100 'meta-modules' representing robust patterns of gene activity in human DFC (**Fig. 2.5b**). Each meta-module represented on average ~1% of all modules and ~1% of all module

genes (**Fig. 2.5b**). To visualize the expression patterns of meta-modules, we formed 'eigenmodules' by calculating PC1 of the REI projection patterns for merged modules in each snRNA-seq dataset. Visualizing these eigenmodules as heat maps revealed highly similar patterns of gene activity for meta-modules in both DFC snRNA-seq datasets (**Fig. 2.5c,d**). We also performed consensus clustering of cell types using CoPA projection patterns (i.e., the transpose of module clustering) (**Fig. 2.5a**), which recapitulated known relationships among subclasses (**Fig. 2.5c,d**). A parallel analysis of two MTG snRNA-seq datasets[29,30] produced strikingly similar results (**Fig. S2.8**).

By calculating the number of modules and genes associated with each meta-module and merged meta-modules (barplots and branch labels in **Fig. 2.5b**), we quantified the frequencies of expression patterns associated with cellular identities at different resolutions in human DFC (**Fig. 2.5b-d**). The initial bifurcation in meta-module clustering revealed that approximately half of all meta-modules have higher relative expression in neurons and half in non-neurons. On average, among non-neurons, vascular cells (endothelial + VLMC) had the highest relative expression for 17.5% of all meta-modules, oligodendrocytes for 16%, astrocytes for 15%, microglia for 11.5%, and OPCs for 2.5%. Among excitatory neurons, L5 ET neurons had the highest relative expression for 11% of all meta-modules, followed by 4% for L6 IT Car3, 3.5% for L2/3 IT, 3.5% for L5/6 NP, 2.5% for L4 IT, 1% for L5 IT, 1% for L6 IT, 0.5% for L6b, and 0% for L6 CT. Among inhibitory neurons, PVALB neurons had the highest relative expression for 3.5% of all meta-modules, followed by 2.5% for SNCG, 1.5% for SST, 1% for Chandelier, 1% for LAMP5, 0.5% for LAMP5 LHX6, 0.5% for SST CHODL, and 0% for PAX6 and VIP. Analysis of MTG produced very similar results (**Fig. S2.8**).

dCoPA reveals coordinated and reproducible dysregulation of modular gene activity in neuropathologies

The consistency of modular gene expression patterns among neocortical subclasses (**Fig. 2.4s**) provides a novel framework for identifying transcriptional perturbations associated with disease. Specifically, we reasoned that comparing CoPA projection patterns for aggregated subclass nuclei derived from normal and pathological human brain samples could reveal cell-type-specific dysregulation of modular gene activity. By comparing the expression levels of coexpressed genes in all nuclei from each subclass, this strategy retains the statistical power of bulk coexpression analysis while mitigating the noise and sparsity inherent to snRNA-seq data and preserving cell-type-specific information. We therefore developed an empirical approach to identify bulk coexpression modules whose expression levels differed significantly in one or more neocortical subclasses from normal and pathological samples. We imposed two criteria: i) the Euclidean distance between the mean expression levels of module genes in case vs. control subclass nuclei must be greater than expected by chance, and ii) all genes in the module must change expression in the same direction (**Fig. 2.6a**).

We evaluated our approach using two large snRNA-seq datasets[29,49] comprising DFC and MTG nuclei from neurotypical adults (CTRL) and individuals with Alzheimer's disease (AD). Combined, these studies profiled 629,013 DFC nuclei / 318,994 MTG nuclei from 64 CTRL subjects and 1,478,438 DFC nuclei / 1,273,567 MTG nuclei from 126 AD subjects. We performed dCoPA separately for each study and brain region using real modules (n=1,016; **Fig. 2.4**) or random modules (i.e., modules formed by randomly selecting genes to match real module sizes). Importantly, after correcting for multiple comparisons, no random modules were significant in any analysis (**Fig. S2.9**).

In contrast, over a third of real modules were significant for at least one cell type in each analysis, with a mean of about two significant cell types per significant module (**Fig. S2.9**).

To ensure the robustness of our findings, we stipulated that significant modules must also be reproduced in both snRNA-seq datasets. More specifically, for each brain region, we compared dCoPA results and retained only those significant modules for which all genes moved in the same direction in the same cell type in both snRNA-seq datasets. Snapshots of two such modules are shown in **Fig. 2.6b-i**. Module 77 was significantly and reproducibly downregulated in L4 IT excitatory neurons from AD samples (**Fig. 2.6f,h**), while module 139 was significantly and reproducibly downregulated in L5 ET and L6 IT excitatory neurons from AD samples (**Fig. 2.6g,i**). We note that the top gene in this module (*TMEM106B*, **Fig. 2.6c**) has been reproducibly implicated in AD pathology via multiple genetic association studies.[50-52] Importantly, although all 70 genes in both modules were consistently and reproducibly downregulated in the same neuronal subclasses, none of these genes were significant by DE analysis of either snRNA-seq dataset using either method.

To summarize and compare dCoPA results between snRNA-seq datasets (Gabitto et al.[29] and Liu et al.[49]) and brain regions (MTG and DFC), we visualized the number of significant modules for each cell type as dot plots and separated modules that were significantly downregulated in AD nuclei from those that were significantly upregulated (**Fig. 2.7a,b** & **Fig. S2.10a,b**). For example, ‘CTRL modules | Gabitto SN’ (**Fig. 2.7a**) depicts the number of bulk coexpression modules from our integrated control cohort that were significantly downregulated in AD MTG nuclei vs. CTRL MTG nuclei for

each cell type using snRNA-seq data from Gabitto et al. Interestingly, the overwhelming majority of significant modules identified by dCoPA were downregulated in AD nuclei (primarily in neurons), while the smaller number of significant modules that were upregulated in AD nuclei were more likely to occur in non-neurons (**Fig. 2.7a,b** & **Fig. S2.10a,b**).

In principle, the bulk coexpression modules used as input for dCoPA can be derived from normal or pathological cohorts. We therefore reran CoPA to identify bulk coexpression modules in the ROSMAP cohort of AD patient samples (n = 248 from DFC).[53] This analysis identified 1,010 modules, which were projected onto AD and CTRL nuclei from Gabitto et al.[29] or Liu et al.[49] in the same fashion as the bulk CTRL modules (**Fig. 2.7a,b** & **Fig. S2.10a,b**). We next compared the number of bulk CTRL or bulk AD modules that were significantly and reproducibly downregulated in each cell type (bottom two rows in **Fig. 2.7a,b**). Even though the bulk CTRL and bulk AD modules were independently derived, the overall pattern of cell-type-specific downregulation identified by dCoPA was extremely reproducible (**Fig. 2.7a,b**: r = 0.93 for MTG and r = 0.96 for DFC), with the greatest number of downregulated modules observed in deep layer intratelencephalic excitatory neurons and PVALB or LAMP5 inhibitory neurons from both regions. Users can also interact with the dot plots in **Fig. 2.7a,b** & **Fig. S2.10a,b** on the CoPA Cabana web site (https://oldhamlab.shinyapps.io/copacabana/), where clicking on a dot reveals the snapshots of modules that are significantly and reproducibly dysregulated in a given cell type.

We next compared the identities of genes comprising the bulk CTRL and bulk AD modules that were significantly and reproducibly downregulated in each cell type. We

observed the most significant overlap (P < 1.0E-14) of these genes in L4 IT, L5 ET, L5 IT, L6 IT, L6 IT Car3, L6b, PVALB, and LAMP5 neurons, with many of the same genes downregulated in multiple neuronal cell types (**Fig. S2.11**). To identify the most reproducible set of affected genes, we took the intersection of all genes comprising bulk CTRL or bulk AD modules that were significantly and reproducibly downregulated in each cell type (bottom two rows in **Fig. 2.7a,b**). We then evaluated how these genes were distributed among cell types and identified deep layer intratelencephalic excitatory neurons and PVALB or LAMP5 inhibitory neurons as the most affected, albeit with some variability between brain regions (**Fig. 2.7c,d**). Enrichment analysis revealed significant over-representation of Gene Ontology[54] categories related to ribosomal function, protein localization, protein conjugation, and protein ubiquitination among module genes that were coordinately and reproducibly downregulated in specific neuronal subtypes (**Fig. 2.7e,f**: left). We also performed an AI-powered multi-agent search of the biomedical literature to identify studies that have implicated these genes in AD biology (**Fig. 2.7e,f**: right).

To evaluate dCoPA for a different type of neuropathology, we obtained additional bulk and snRNA-seq data from patients with schizophrenia (SCZ) and controls. We analyzed the same set of bulk CTRL modules as we did for the AD analysis (n = 1,016), while bulk SCZ modules (n = 1,083) were identified by applying CoPA to 171 DFC samples from SCZ patients from the Brainseq Phase 1 study.[42] Using dCoPA, we analyzed the CTRL and SCZ modules in two snRNA-seq datasets that collectively profiled 461,353 nuclei from 77 CTRL subjects and 385,547 nuclei from 71 SCZ subjects.[55] As with AD, the vast majority of significant and reproducible modules

identified by dCoPA were downregulated in SCZ neurons, while the few that were upregulated in SCZ were primarily observed in non-neurons (**Fig. 2.8a,b**). We then compared the number of bulk CTRL or bulk SCZ modules that were significantly and reproducibly downregulated in each cell type and again found that the overall pattern of cell-type-specific downregulation identified by dCoPA was highly reproducible (r = 0.78; bottom two rows in **Fig. 2.8a,b**).

We next compared the identities of genes comprising the bulk CTRL and bulk SCZ modules that were significantly and reproducibly downregulated in each cell type. We observed the most significant overlap (P < 1.0E-7) of these genes in L5 IT, L6 IT, L6 IT Car3, L6b, SST, and VIP neurons, with many of the same genes downregulated in multiple neuronal cell types (**Fig. 2.8c**). To identify the most reproducible set of affected genes, we took the intersection of all genes comprising bulk CTRL or bulk SCZ modules that were significantly and reproducibly downregulated in each cell type (bottom two rows in **Fig. 2.8a**). We then evaluated how these genes were distributed among cell types and identified SST, VIP, and L6b neurons as the most affected (**Fig. 2.8d**). Enrichment analysis revealed significant over-representation of Gene Ontology biological processes related to protein targeting to membrane and microtubules among module genes that were coordinately and reproducibly downregulated in these neuronal subtypes (**Fig. 2.8e** [left]). We also performed an AI-powered multi-agent search of the biomedical literature to identify studies that have implicated these genes in SCZ biology (**Fig. 2.8e**: right). An example of one module with significant and reproducible downregulation in multiple SCZ excitatory neuron subtypes is shown in **Fig. 2.8f**, with similar visualizations for all significant modules available in CoPA Cabana

(https://oldhamlab.shinyapps.io/copacabana/). Interestingly, we noticed that some modules identified by dCoPA were significant for SCZ and AD. Specifically, we identified 18 examples of modules that were significantly, coordinately, and reproducibly downregulated in the same cell type(s) in AD and SCZ. This overlap was highly significant (P = 3.4e-24).

Discussion

CoPA is a novel and statistically motivated approach for combining the power of bulk tissue sampling with the precision of single-cell analysis to reveal the origins of highly reproducible gene coexpression modules and their relative importance among cell types. By applying CoPA to vast amounts of public gene expression data, we have shown that very few gene coexpression modules in human neocortex are cell-type-specific; however, subtle differences in modular gene activity among cell types are nevertheless extremely reproducible. This reproducibility enables comparisons of module projection patterns between single-cell datasets derived from normal and pathological human tissue samples with dCoPA. Using this approach, we identified gene coexpression modules that are uniformly and reproducibly dysregulated in specific neocortical cell types from patients with AD or SCZ. To share our findings, we have created a user-friendly R Shiny app called CoPA Cabana (https://oldhamlab.shinyapps.io/copacabana/) that can also perform CoPA for hundreds of thousands of gene coexpression modules in OMICON (https://theomicon.ucsf.edu), which are derived from standardized analysis of 116 datasets representing >17K normal and neoplastic human brain samples.

Our central insight is that the population structure of genomic activity – efficiently discovered through covariation analysis of intact tissue samples – provides a natural framework for interpreting single-cell datasets. Because large bulk datasets may represent thousands of individuals and billions of cells, they provide enormous statistical power to discern highly reproducible modules of genomic activity, which also reduce the dimensionality of gene expression data. By projecting these modules onto pseudobulked cell types from single-cell datasets, CoPA clarifies the relative importance of these modules for each cell type, and also how cell types differ from one another. Given that neocortical subclass relationships were faithfully reconstructed by clustering bulk coexpression module projection patterns (e.g., **Fig. 2.5c**), it follows that bulk coexpression modules could potentially be used to refine cell-type annotations in neocortical single-cell datasets. However, any such effort should acknowledge that similarities far outweigh differences in gene expression among neuronal cell types in human neocortex. For example, the average genome-wide correlation among all neuronal subclasses (excitatory and inhibitory) in human DFC and MTG is ~0.9; for oligodendrocytes and OPCs, the corresponding value is ~0.7. Therefore, claims of ever more granular neuronal subtypes should be treated with caution unless they have been unambiguously replicated in multiple, independent datasets.

dCoPA provides a statistical framework for comparing CoPA projection patterns in single-cell datasets from different cohorts. Unlike standard DE analysis, which treats genes separately and ignores their coexpression relationships, dCoPA uses the covariation structure of genomic activity in bulk datasets as a scaffold for inquiry. By comparing the mean expression levels of coexpressed genes between vast numbers of

cells or nuclei for each cell type and requiring that expression differences between cohorts are significant, uniform, and reproducible, we introduce a novel and robust strategy for revealing pathological gene activity in single-cell datasets. Application of this strategy to AD and SCZ produced several interesting findings.

First, in both AD and SCZ, dCoPA identified many more modules that were significantly downregulated than upregulated. Second, most of the downregulated modules implicated neurons, while most of the upregulated modules implicated non-neurons. Third, many significant modules were coordinately downregulated in multiple neuronal subclasses (often including excitatory and inhibitory subclasses). Fourth, there were more significant dCoPA modules in AD MTG than AD DFC, consistent with known AD pathology.[56] Fifth, implicated cell types were highly consistent whether bulk coexpression modules were derived from normal or pathological cohorts. And sixth, although the most affected cell types differed by pathology (**AD:** L6 IT, L4 IT, PVALB, LAMP5; **SCZ:** SST, VIP, L6b), a significant number of dCoPA modules were dysregulated in the same direction and the same cell type(s) in AD and SCZ.

Collectively, our findings highlight distinct gene coexpression modules that are selectively vulnerable to dysregulation in AD and SCZ. Hardly any of these modules are cell-type-specific, and most of the genes comprising them would not be identified by standard DE analysis of individual cell types. However, these genes paint a detailed portrait of the cellular activities most impacted by disease. For example, the genes that are most recurrently identified by dCoPA for ADsuggest a coordinated neuronal collapse of protein synthesis, membrane trafficking / lysosomal clearance, axonal / synaptic delivery, and nuclear gene regulatory control. In SCZ, the same analysispoints to

reduced expression of genes involved in synaptic structure, vesicle cycling, receptor localization, and dendritic spine structure. Our results also raise the tantalizing possibility that targeted reversal of dysregulated modules identified by dCoPA may restore normal cellular functions and ameliorate pathological symptoms.

Our study has several important limitations. First, CoPA projections can be misleading if bulk coexpression modules are driven by cell types or states that are not represented in single-cell datasets. For example, none of the snRNA-seq datasets we analyzed included ependymal or choroid plexus cells. Second, the accuracy of module projections may be limited by the sample size available for each annotated cell type; in some snRNA-seq datasets, certain cell types were represented by < 200 nuclei (e.g., VLMC and SST CHODL), limiting power. Third, there are many ways to define gene coexpression modules in bulk datasets. We took a conservative approach and iteratively removed all groups of highly correlated genes without merging them, which means that some modules may be closely related and functionally redundant. However, dCoPA analysis of modules derived independently from normal or pathological bulk datasets revealed remarkable consistency in the patterns of affected cell types, underscoring the robustness of our approach.

In summary, we have described a novel analytical strategy that combines the power of bulk coexpression analysis with the precision of single-cell methods to reveal the cellular origins of highly reproducible modules of genomic activity in health and disease. We have also created the CoPA Cabana web site (https://oldhamlab.shinyapps.io/copacabana/), which accompanies this study and provides a novel resource for the neuroscientific community by hosting all CoPA and

dCoPA results from our analyses. We expect that application of CoPA and dCoPA to other pathological conditions and omics dataset types will clarify the organization of genomic activity in disparate tissues and the combinations of molecular programs and cell types that are selectively impacted by disease.

Figures

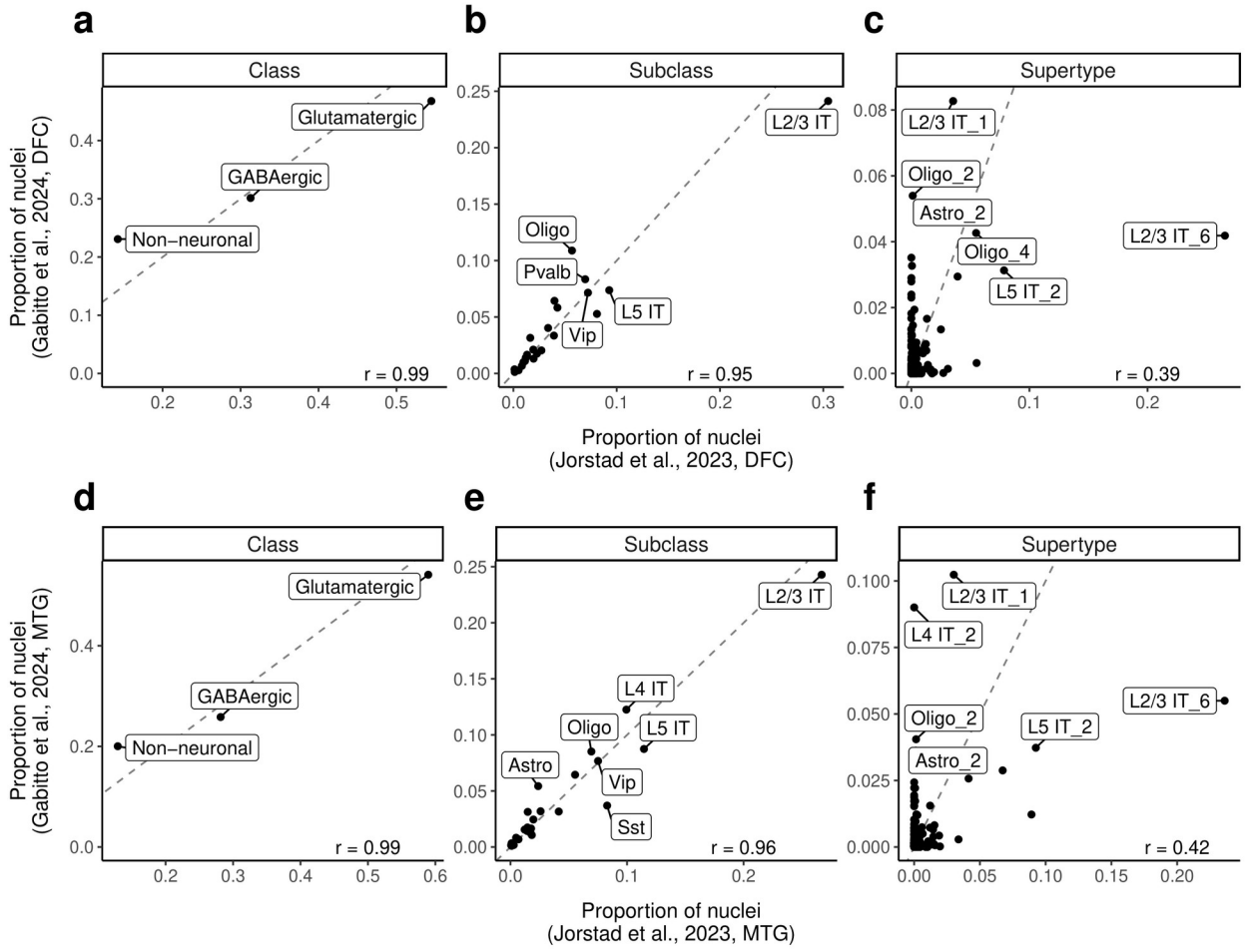
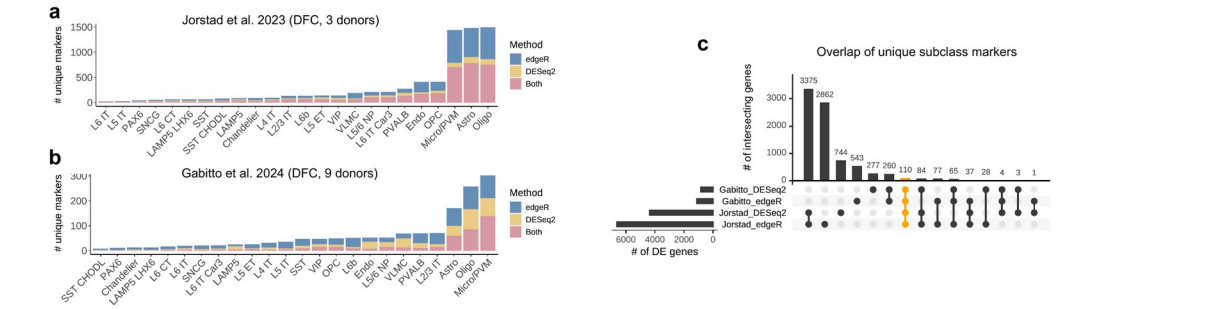


Figure S2.1: Cell class and subclass proportions are more highly conserved than supertype proportions in human neocortex.

a-f) Proportion of nuclei in adult human dorsal prefrontal cortex (DFC: **a-c**) or middle temporal gyrus (MTG: **d-f**) assigned to each cell class (**a, d**), subclass (**b, e**), or supertype (**c, f**) in two snRNA-seq datasets.[29,30]



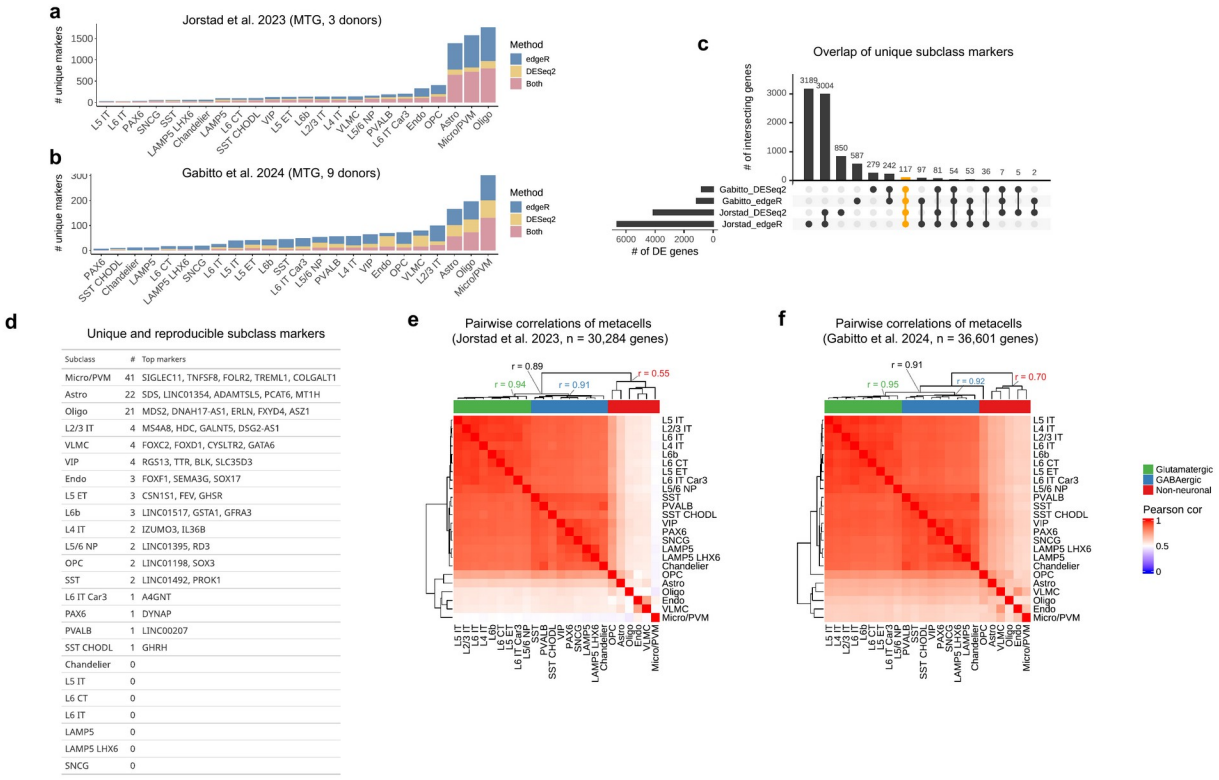
d

Unique and reproducible subclass markers

Subclass	#	Top markers
Micro/PVM	37	MPEG1, TNFSF8, RNF166, CLEC17A, SERPINA1
Astro	21	PHKG1, FGFRL1, GRTP1-AS1, LINC01354, MT1H
Oligo	21	PIP, NDUFS7, DNAH17-AS1, C11orf71, ERLN
L5/6 NP	5	CD200R1L, OXT, RFPL4B, ZP2, LINC01307
SST	4	NUGGC, SLC9A4, PROK1, LINC00308
Endo	3	CRIPTO, CHMP4C, BTNL3
L2/3 IT	3	MS4A8, WFDC10A, SSMEM1
L6b	3	RSPO1, SACK1B, MS4A2
VLMC	3	AGTR1, NKX6-1, FZD10
L5 ET	2	CSN1S1, PRR27
OPC	2	PRRX2, SOX3
PVALB	2	PRAMEF20, CLEC4M
VIP	2	LEFTY2, SLC35D3
L6 CT	1	SLC13A2
PAX6	1	DYNAP
Chandelier	0	
L4 IT	0	
L5 IT	0	
L6 IT	0	
L6 IT Car3	0	
LAMP5	0	
LAMP5 LHX6	0	
SNCG	0	
SST CHODL	0	

Figure 2.1: Cell type marker genes in human frontal cortex vary by choice of study and algorithm.

a-b) Number of unique marker genes (FDR < .05) identified by differential expression (DE) analysis of snRNA-seq data from two studies[29,30] of normal adult human dorsal frontal cortex (DFC) using edgeR,[33] DESeq2,[34] or both. **c)** Overlap of unique marker genes from **a** & **b** for all cell types. **d**) Unique and reproducible marker genes for each cell type ranked by significance. **e-f**) Heatmaps of genome-wide correlations among pseudobulked cell types from Jorstad et al.[30] (**e**) and Gabitto et al.[29] (**f**). Mean correlations among metacells of a given class (glutamatergic, GABAergic, neuronal, non-neuronal) are reported at dendrogram branch points.



Subclass	#	Top markers
Micro/PVM	41	SIGLEC11, TNFSF8, FOLR2, TREML1, COLGALT1
Astro	22	SDS, LINC01354, ADAMTSL5, PCAT6, MT1H
Oligo	21	MDS2, DNAH17-AS1, ERLN, FXYD4, ASZ1
L2/3 IT	4	MS4A8, HDC, GALNT5, DSG2-AS1
VLMC	4	FOXC2, FOXD1, CYSLTR2, GATA6
VIP	4	RGS13, TTR, BLK, SLC35D3
Endo	3	FOXF1, SEMA3G, SOX17
L5 ET	3	CSN1S1, FEV, GHSR
L6b	3	LINC01517, GSTA1, GFRA3
L4 IT	2	IZUMO3, IL36B
L5/6 NP	2	LINC01395, RD3
OPC	2	LINC01198, SOX3
SST	2	LINC01492, PROK1
L6 IT Car3	1	A4GNT
PAX6	1	DYNAP
PVALB	1	LINC00207
SST CHODL	1	GHRH
Chandelier	0	
L5 IT	0	
L6 CT	0	
L6 IT	0	
LAMP5	0	
LAMP5 LHX6	0	
SNCG	0	

Figure S2.2: Cell type marker genes in human temporal cortex vary by choice of study and algorithm.

a-b) Number of unique marker genes (FDR < .05) identified by differential expression (DE) analysis of snRNA-seq data from two studies[29,30] of normal adult human middle temporal gyrus (MTG) using edgeR,[33] DESeq2,[34] or both. **c)** Overlap of unique marker genes from **a** & **b** for all cell types. **d**) Unique and reproducible marker genes for each cell type ranked by significance. **e-f**) Heatmaps of genome-wide correlations among pseudobulked cell types from Jorstad et al.[30] (**e**) and Gabitto et al.[29] (**f**). Mean correlations among metacells of a given class (glutamatergic, GABAergic, neuronal, non-neuronal) are reported at dendrogram branch points.

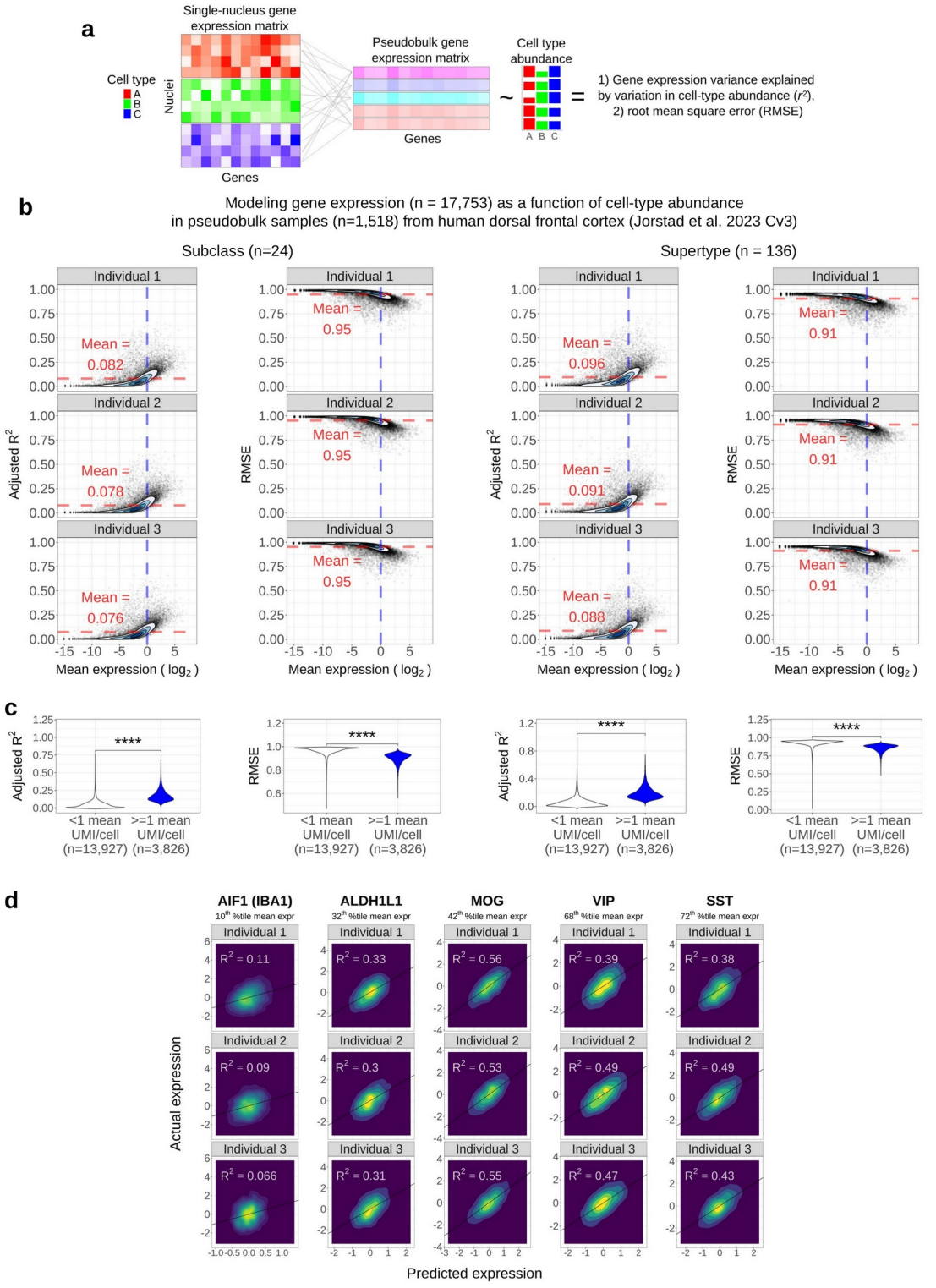


Figure 2.2: Genome-wide expression variation in pseudobulked snRNA-seq data from human frontal cortex is poorly explained by cell-type abundance, particularly for low-expressed genes.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a) Schematic of pseudobulk modeling approach. Pseudobulk samples (n=1,518) were created by randomly sampling 10% of Cv3 DFC nuclei from each donor in Jorstad et al.,[30] recording their identities, and summing expression levels for each gene. Each gene was then used as the response variable in a multiple linear regression model with the abundance of each subclass (left) or supertype (right) as predictors. **b**) Adjusted R^2 and root mean square error (RMSE) modeling results for all genes vs. mean expression level (log_2). Red dotted lines denote mean modeling performance, while blue dotted lines denote mean expression of 1 UMI/nucleus.[19,20] **c**) Violin plots of modeling performance from (**b**), stratified by genes below and above this critical threshold. ****P < 0.0001. **d**) Predicted (fitted) vs. actual expression of select marker genes with different expression levels.

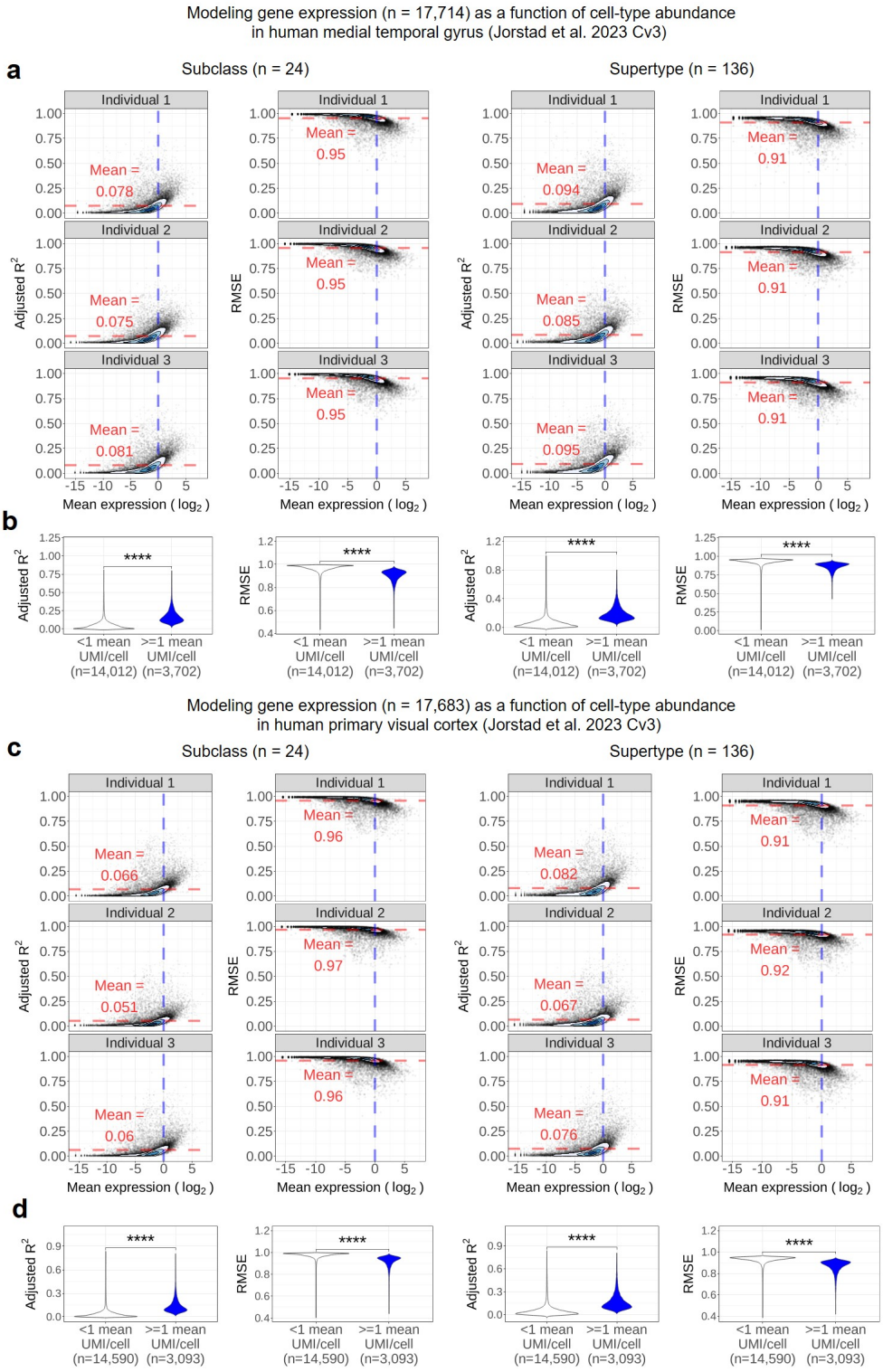


Figure S2.3: Genome-wide expression variation in pseudobulked snRNA-seq data from human temporal and visual cortex is poorly explained by cell-type abundance, particularly for low-expressed genes.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a-d) Variation in cell type abundance was used to predict genome-wide expression levels in pseudobulked snRNA-seq data[30] from middle temporal gyrus (**a**, **b**) or primary visual cortex (**c**, **d**) via multiple linear regression as described in **Fig. 2.2.** (**a**, **c**) Adjusted R^2 and root mean square error (RMSE) modeling results for all genes vs. mean expression level (log_2). Red dotted lines denote mean modeling performance, while blue dotted lines denote mean expression of 1 UMI/nucleus.[19,20] (**b**, **d**) Violin plots of modeling performance from (**a**, **c**), stratified by genes below and above this critical threshold. ****P < 0.0001

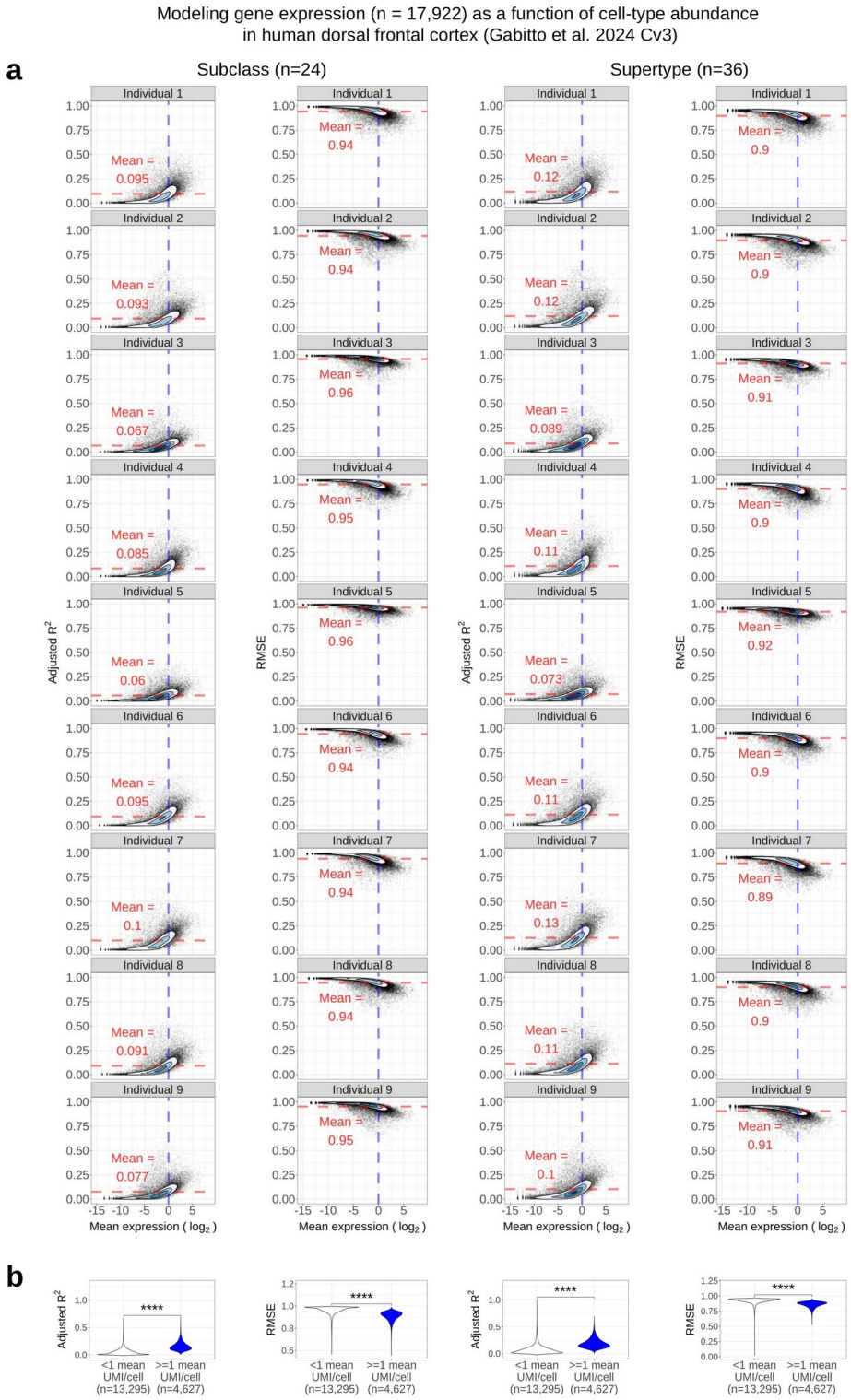


Figure S2.4: Genome-wide expression variation in pseudobulked snRNA-seq data from human frontal cortex is poorly explained by cell-type abundance, particularly for low-expressed genes.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a-b) Variation in cell type abundance was used to predict genome-wide expression levels in pseudobulked snRNA-seq data[29] from dorsal frontal cortex via multiple linear regression as described in **Fig. 2.2. a**) Adjusted R^2 and root mean square error (RMSE) modeling results for all genes vs. mean expression level (log_2). Red dotted lines denote mean modeling performance, while blue dotted lines denote mean expression of 1 UMI/nucleus.[19,20] (**b**) Violin plots of modeling performance from (**a**), stratified by genes below and above this critical threshold. ****P < 0.0001

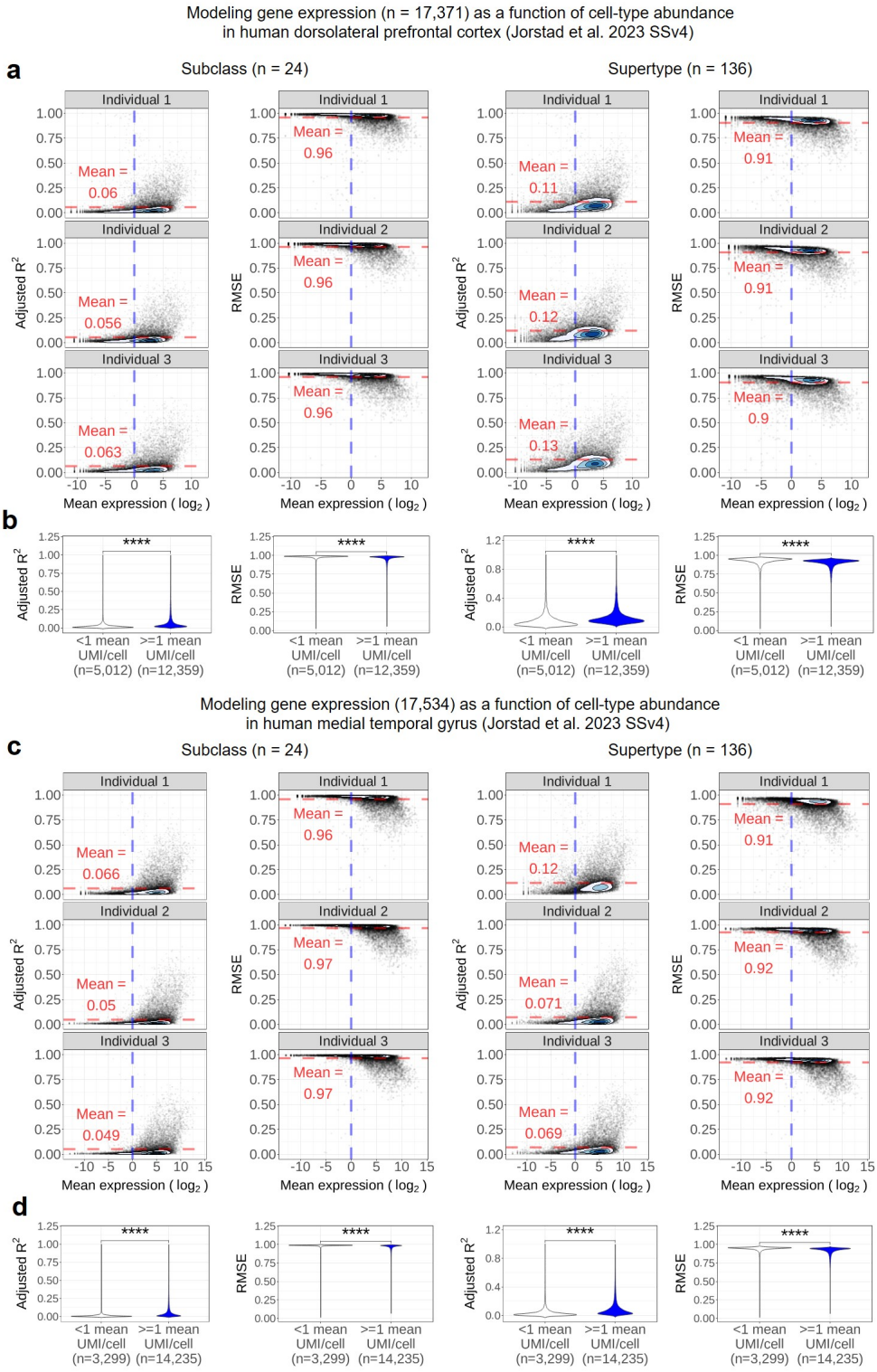


Figure S2.5: Genome-wide expression variation in pseudobulked snRNA-seq data from human neocortex is poorly explained by cell-type abundance, even with deeper sequencing.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a-d) Variation in cell type abundance was used to predict genome-wide expression levels in pseudobulked snRNA-seq data[30] from frontal cortex (**a**, **b**) or temporal cortex (**c**, **d**) via multiple linear regression as described in **Fig. 2.2.** (**a**, **c**) Adjusted R^2 and root mean square error (RMSE) modeling results for all genes vs. mean expression level (log_2). Red dotted lines denote mean modeling performance, while blue dotted lines denote mean expression of 1 UMI/nucleus.[19,20] (**b**, **d**) Violin plots of modeling performance from (**a**, **c**), stratified by genes below and above this critical threshold. ****P < 0.0001

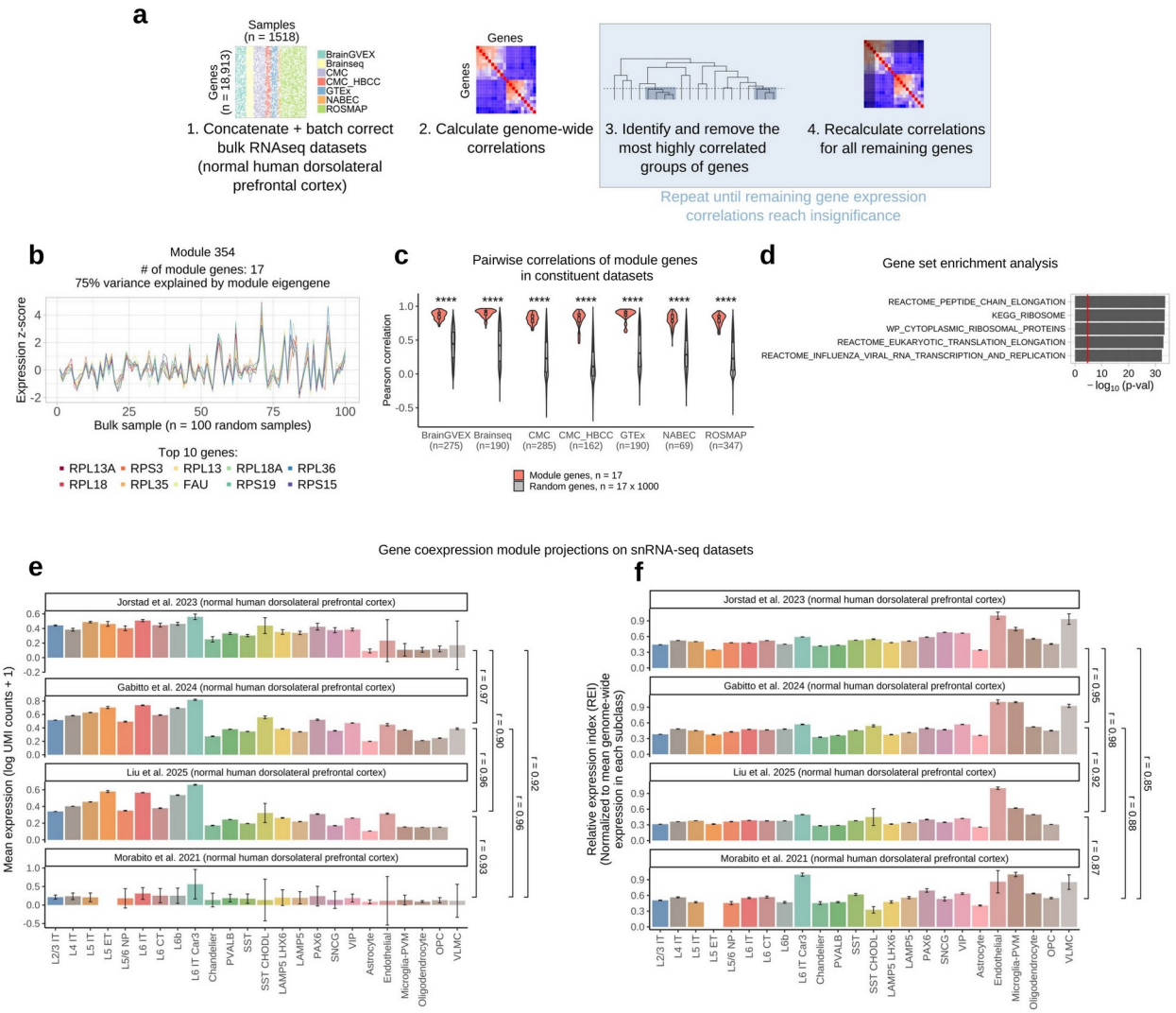


Figure 2.3: Covariation Projection Analysis (CoPA) reveals the cellular origins of bulk gene coexpression modules.

a) Strategy for identifying gene coexpression modules in a large bulk RNAseq dataset comprising neurotypical adult human DFC samples (n=1,518) from six studies.[38-44] **b-f)** 'Snapshot' of one bulk gene coexpression module (Module 354). **b**) Expression patterns (z-scores) for module genes (n = 17) in 100 random DFC samples; 'module eigengene' is PC1 of the gene coexpression module.[47] **c**) Pairwise correlations of module genes in each of the input datasets referenced in (**a**) vs. pairwise correlations of random genes. **d**) Top results from enrichment analysis (one-sided Fisher's exact test) of module genes using a large collection of gene sets from the Molecular Signatures Database[48] and OMICON (https://theomicon.ucsf.edu). **e-f**) Module projections onto pseudobulked cell types from four snRNA-seq studies of human DFC.[29-31,49] Projections were calculated by averaging the expression of all module genes (natural log of UMI counts + 1) over all cells from each cell type in each dataset (**e**). A

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
relative expression index (REI) was calculated by dividing values in (**e**) by the mean expression of all genes for each subclass, then scaling values to a maximum of 1 (**f**). Pearson correlations of projections are shown at right.

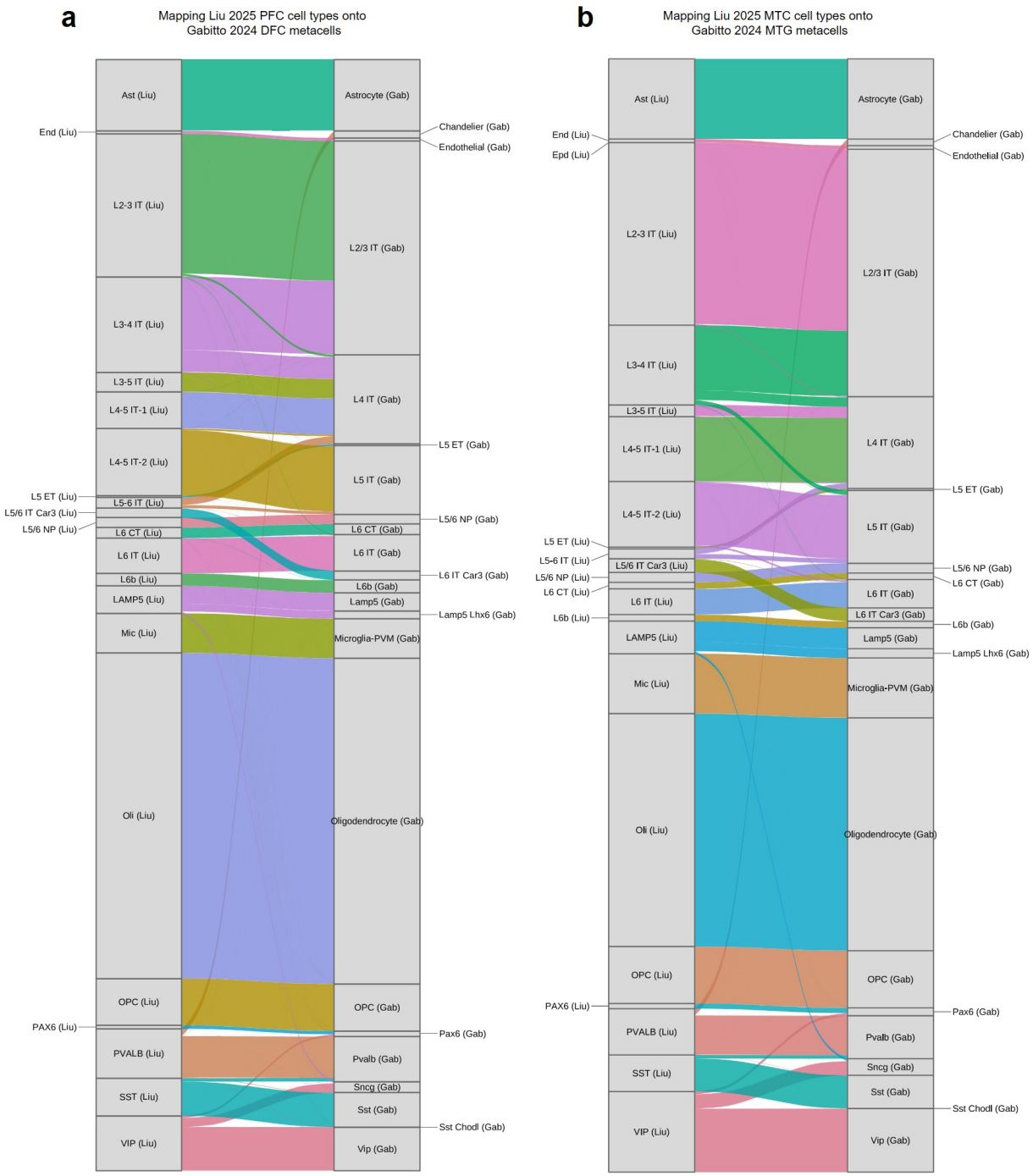


Figure S2.6: Liu et al. nuclei map cleanly to subclass identities from Gabitto et al.

a-b) River plots display mapping of human frontal cortex (**a**) and temporal cortex (**b**) nuclei from Liu et al.[49] to subclass identities from Gabitto et al.[29]

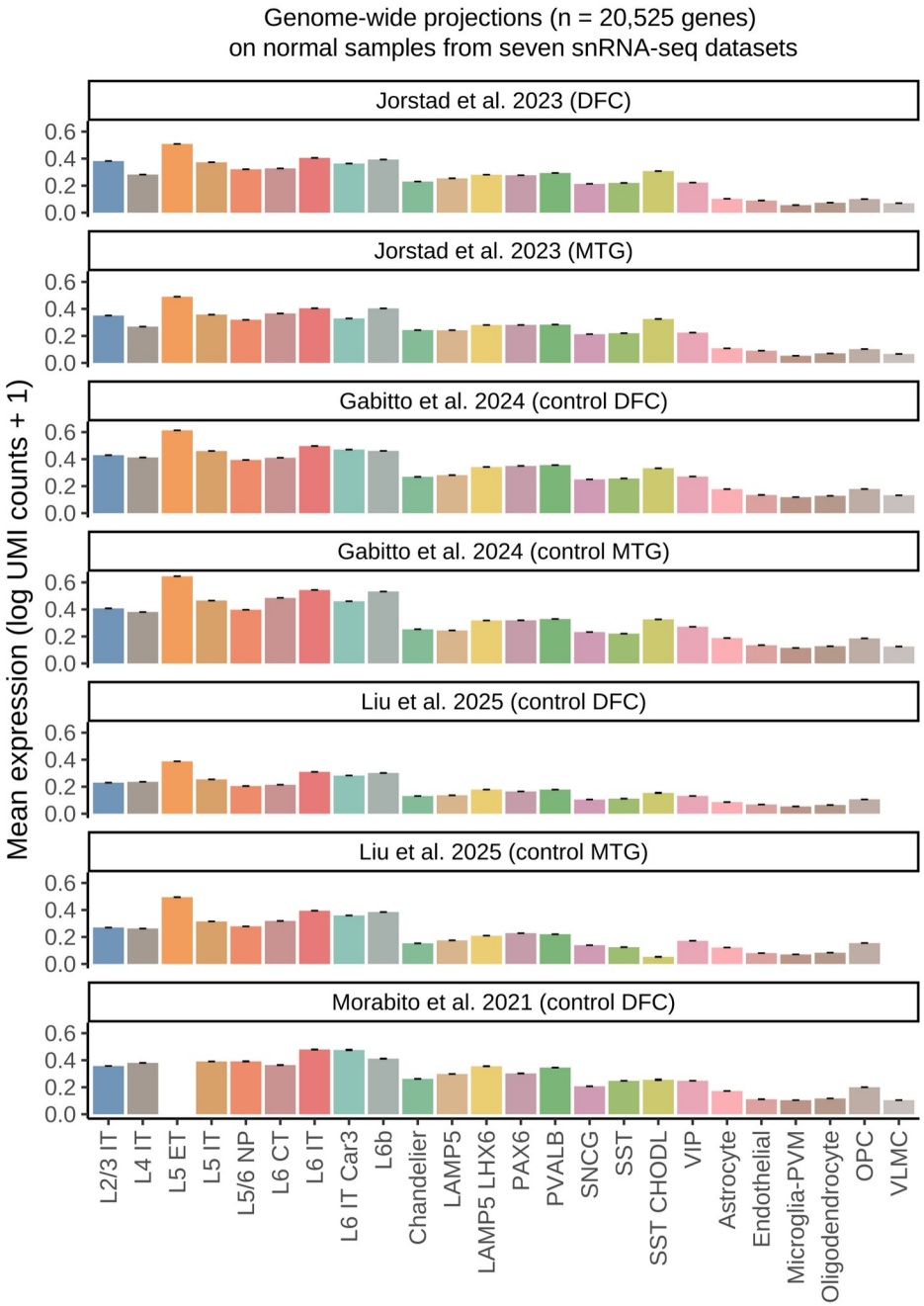


Figure S2.7: Genome-wide expression levels vary significantly among human neocortical subclasses.

Expression levels for all genes (n = 18,913) were averaged for each cell type using snRNA-seq data from four studies[29-31,49] and two brain regions. DFC = dorsal frontal cortex; MTG = middle temporal gyrus.

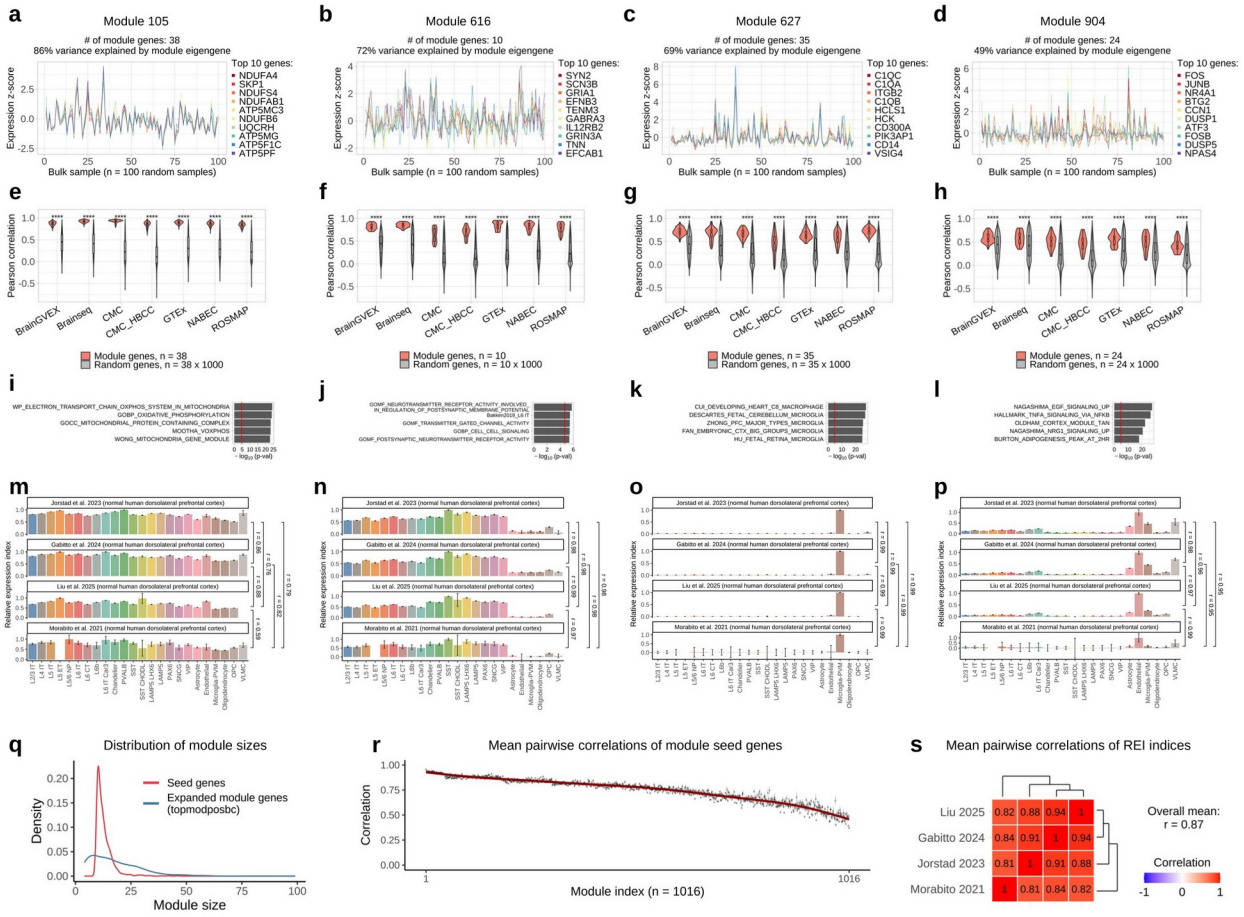


Figure 2.4: Bulk gene coexpression modules in human neocortex reflect highly reproducible patterns of genomic activity among neocortical cell types.

a-p) Snapshots of four representative gene coexpression modules**. a-d**) Expression patterns (z-scores) for module genes in 100 random DFC samples. **e-h**) Pairwise correlations of module genes in each of the input datasets (**Fig. 2.3a**) vs. pairwise correlations of random genes. **i-l**) Top results from enrichment analysis (one-sided Fisher's exact test) of module genes using a large collection of gene sets from the Molecular Signatures Database[48] and OMICON (https://theomicon.ucsf.edu). **m-p**) Module projections onto pseudobulked cell types from four snRNA-seq studies of human DFC,[29-31,49] with relative expression indices calculated as described in **Fig. 2.3**. **q**) Sizes for all modules (n=1,016) using two module definitions (seed: initial set of module genes identified by the algorithm in **Fig. 2.3a** and used to form the module eigengene; topmodposbc: all genes that were positively and maximally correlated with the module eigengene below a Bonferroni-corrected significance threshold). **r**) Median pairwise correlations of seed genes for all modules. Vertical bars: +/- two S.E.; red line: smoothing spline. **s**) Mean pairwise correlations of projection indices (REI) from the four snRNA-seq datasets featured in (**m-p**) for all modules.

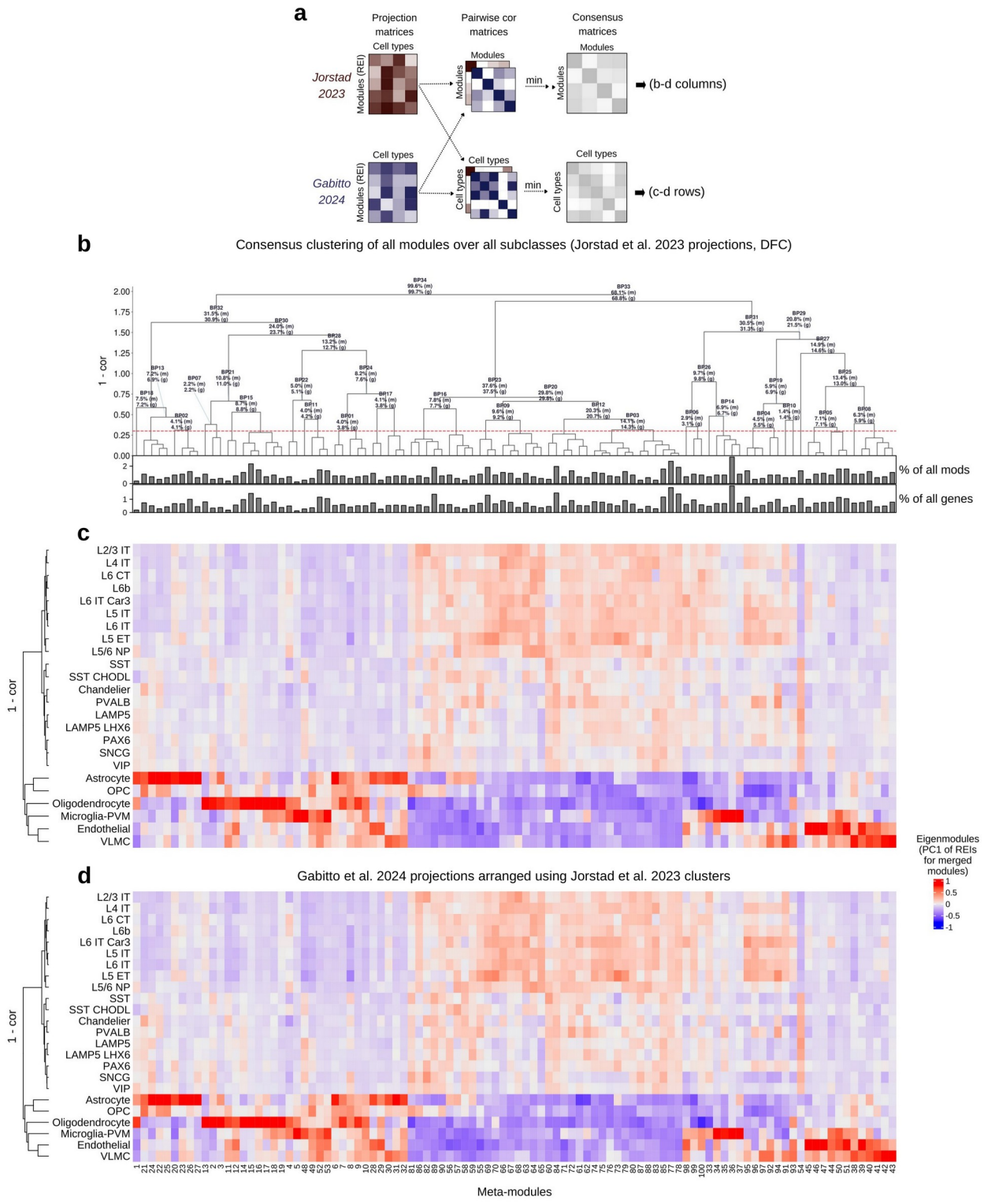


Figure 2.5: Clustering CoPA projection patterns reveals the major motifs of gene activity in human dorsal frontal cortex (DFC) cell types.

a) Schematic of CoPA projection clustering. Left: projection matrices (REI vectors

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
for 1,016 modules and 24 cell types) from two snRNA-seq datasets.[29,30] Middle: pairwise correlation (cor) matrices for modules (top) and cell types (bottom) derived from projection matrices. Right: consensus matrices for modules (top) and cell types (bottom) formed by retaining the minimum cor from both datasets at each position. Consensus cell types and modules were clustered using 1-cor and complete linkage. Meta-modules were identified by cutting the dendrogram at a height equal to the top 5% of pairwise correlations to merge modules (minimum meta-module size: three modules). **b**) Hierarchical clustering of meta-modules (n=100) using DFC snRNA-seq data from Jorstad et al.[30] Barplots below the dendrogram report the percentages of all modules (top) and genes (bottom) comprising each meta-module. Branch points above the red dashed line report the percentages of all constituent modules ('m') and genes ('g'). **c-d**) Heatmaps of eigenmodules (first principal components of REI vectors for merged modules) produced using DFC snRNA-seq data from Jorstad et al.[30] (**c**) and Gabitto et al.[29] (**d**). Cell types were clustered as described in (**a**).

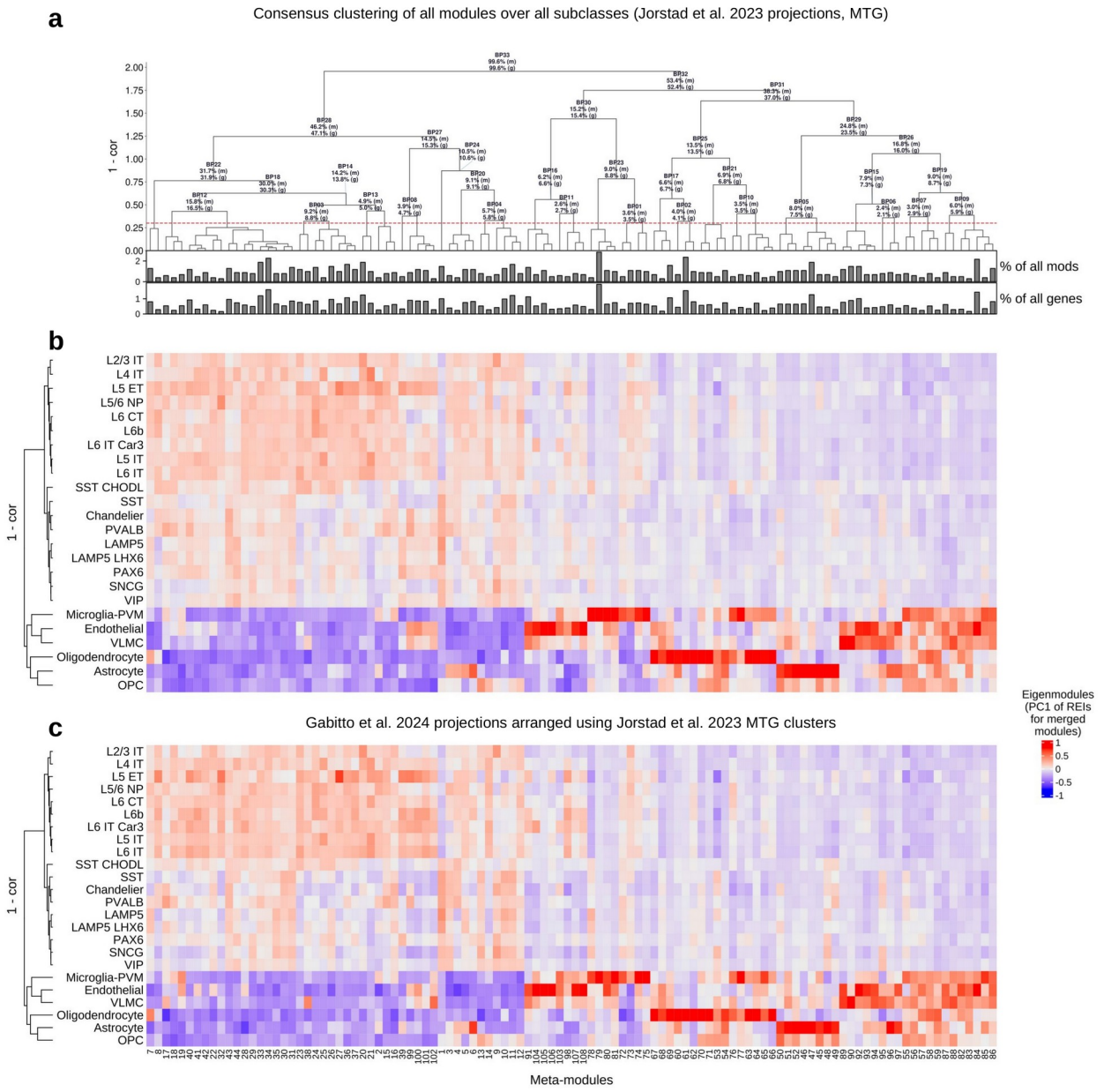


Figure S2.8: Clustering CoPA projection patterns reveals the major motifs of gene activity in human middle temporal gyrus (MTG) cell types.

Clustering was performed as described in **Fig. 2.5a**. **a**) Hierarchical clustering of meta-modules (n=108) using MTG snRNA-seq data from Jorstad et al.[30] Barplots below the dendrogram report the percentages of all modules (top) and genes (bottom) comprising each meta-module. Branch points above the red dashed line report the percentages of all constituent modules ('m') and genes ('g'). **b-c**) Heatmaps of eigenmodules (first principal components of REI vectors for merged modules) produced using MTG snRNA-seq data from Jorstad et al.[30] (**b**) and Gabitto et al.[29] (**c**). Cell types were clustered as described in **Fig. 2.5a**.

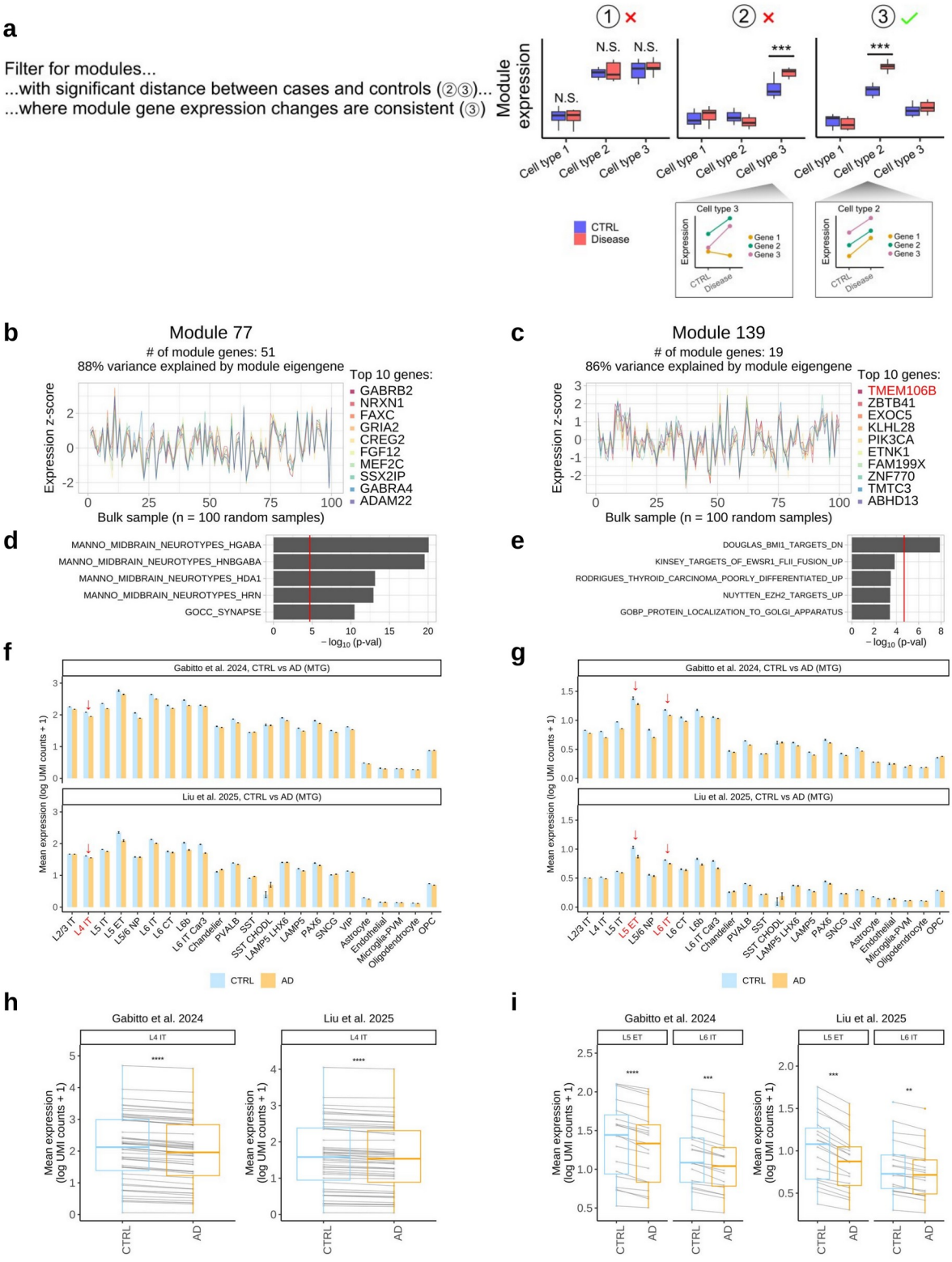


Figure 2.6: Differential CoPA (dCoPA) identifies gene coexpression modules that are significantly and uniformly dysregulated in specific cell types in disease.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
a) Schematic of dCoPA module identification strategy, which uses two criteria: i) the Euclidean distance between the mean expression levels of all module genes (averaged over all subclass nuclei) in cases vs. controls (CTRL) must be greater than expected by chance, and ii) all module genes must change expression in the same direction. **b-i**) Examples of two modules that meet dCoPA filtering criteria for Alzheimer's disease (AD). **b-c**) Expression patterns (z-scores) for module genes in 100 random bulk DFC samples from our input dataset (**Fig. 2.3a**). **d-e**) Top results from enrichment analysis (one-sided Fisher's exact test) of module genes using a large collection of gene sets from the Molecular Signatures Database[48] and OMICON (https://theomicon.ucsf.edu). **f-g**) Projections of module genes onto two snRNA-seq datasets that analyzed CTRL and AD patient samples,[29,49] split by diagnosis. Significant cell types are marked by red arrows. **h-i**) Mean expression levels of all module genes in AD and CTRL nuclei for each cell type marked in (**f-g**). Asterisks indicate the empirical significance of module expression differences between AD and CTRL (compared to 10,000 randomly generated modules of the same size). *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001.

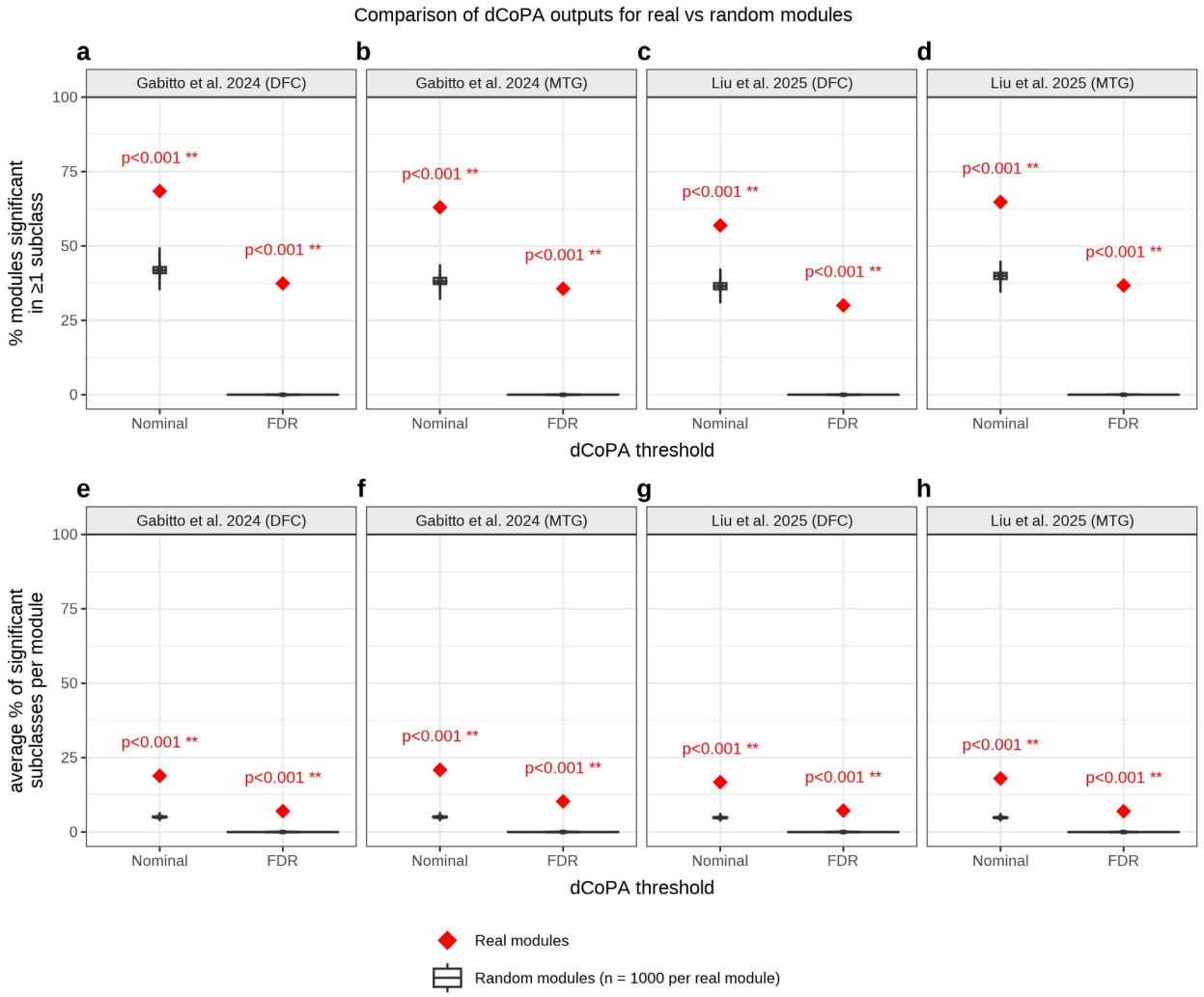
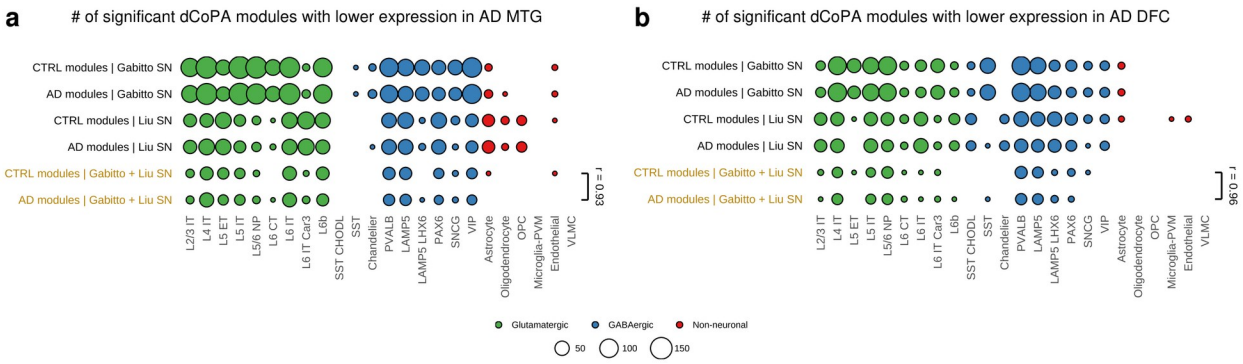
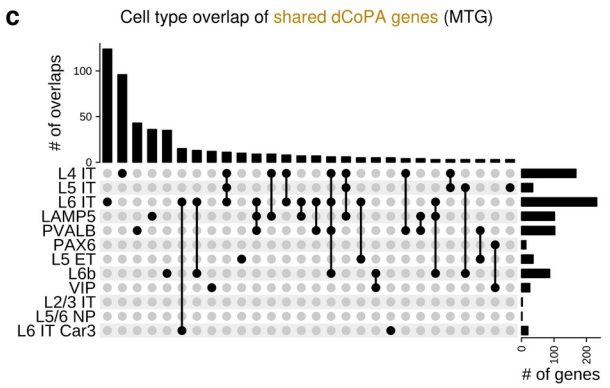
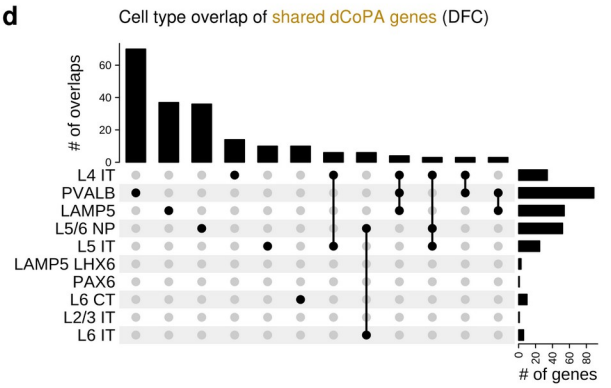


Figure S2.9: Application of dCoPA to random modules produces no significant results.

Comparison of dCoPA results for real modules (n=1,016) and random modules (formed by randomly selecting genes to match real module sizes [n=1,000 random modules per real module]). Four snRNA-seq datasets were analyzed from two studies[29,49] and two brain regions: dorsal frontal cortex (DFC) and middle temporal gyrus (MTG). **a-d)** Percentage of all real (red) or random (black) modules with at least one significant cell type difference. **e-h)** The average percentage of significant subclasses per module. After correcting for multiple comparisons (FDR), no significant cell type differences were identified in random modules.







e Summary of shared dCoPA genes (MTG)

Cell type	Genes (n)	Top 3 GO gene sets (by FDR-adjusted p-value)	AD-associated genes
L2/3 IT	3	—	—
L4 IT	168	1. GOBP_PROTEIN_POLYUBIQUITINATION (padj=4.24e-03) 2. GOMF_UBIQUITIN_LIKE_PROTEIN_TRANSFERASE_ACTIVITY (padj=4.69e-03) 3. GOBP_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_CONJUGATION_OR_REMOV (padj=1.59e-02)	PAFAH1B1, SNAP25, ARCN1, VDAC1, RTN3, YWHAZ, STIM2, MME, CD200, HPRT1
L5 ET	36	—	MAPK8, MME, RORB, AKT3, ATP2A2, SEPTIN6, SLC9A6, ADGRB3, PPP2R2C, PTPRN2
L5 IT	35	—	VDAC1, PDCD6IP, MDH1, CISD1, PCMT1, HACE1, BRWD1, CPNE4, NDFIP1, RIMS1
L5/6 NP	2	—	BRWD1
L6 IT	232	1. GOBP_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICULUM (padj=5.71e-21) 2. GOBP_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICU (padj=5.71e-21) 3. GOBP_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE (padj=7.79e-20)	ARCN1, TIPRL, TMEM106B, OPA1, VPS35, VDAC1, BNIP3, TBK1, RTN3, YWHAZ
L6 IT Car3	20	1. GOBP_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE (padj=1.47e-36) 2. GOCC_CYTOSOLIC_RIBOSOME (padj=1.89e-36) 3. GOBP_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICU (padj=9.20e-36)	RPS29, TOMM7
L6b	87	1. GOBP_CELLULAR_COMPONENT_MORPHOGENESIS (padj=1.46e-03) 2. GOBP_ENDOMEMBRANE_SYSTEM_ORGANIZATION (padj=3.94e-03) 3. GOBP_CELL_PART_MORPHOGENESIS (padj=4.59e-03)	PAFAH1B1, TMEM106B, SV2A, RTN3, SYNJ1, OXR1, MADD, PANX1, ATP2A2, PIKFYVE
LAMP5	102	1. GOBP_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_CONJUGATION_OR_REMOV (padj=2.81e-04) 2. GOBP_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_CONJUGATION (padj=8.23e-04) 3. GOCC_UBIQUITIN_LIGASE_COMPLEX (padj=2.80e-03)	MAPK8, TBK1, RANBP9, PIK3R1, PDCD6IP, AKT3, NRG3, CRBN, RCHY1, ADK
PAX6	14	—	MAPK8, CUL2, SMARCA2
PVALB	103	1. GOBP_CELLULAR_MACROMOLECULE_LOCALIZATION (padj=2.45e-03) 2. GOBP_PROTEIN_LOCALIZATION_TO_NUCLEUS (padj=3.63e-03) 3. GOBP_PROTEIN_IMPORT (padj=8.51e-03)	MTOR, TBK1, RTN3, RANBP9, HPRT1, PIK3R1, DNM1L, MARCHF7, AKT3, CUL3
VIP	26	—	OXR1, USP14, VAPB, G3BP2, CLTC, NALCN, PARG, HS6ST2, ZNF33B

f Summary of shared dCoPA genes (DFC)

Cell type	Genes (n)	Top 3 GO gene sets (by FDR-adjusted p-value)	AD-associated genes
L2/3 IT	1	—	USP44
L4 IT	34	—	DNM1L, LRPPRC, PPP2R2B, N4BP2L1, CEP57, KLHL42, MTMR2, USP44, ZBTB24, ZNF287
L5 IT	25	—	PDCD6IP, PPP2R2B, SLC31A1, UBE4B, N4BP2L1, NRDC, USP44, UBE3B
L5/6 NP	52	1. GOBP_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICU (padj=3.05e-24) 2. GOBP_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE (padj=1.67e-23) 3. GOCC_CYTOSOLIC_RIBOSOME (padj=2.18e-23)	MAPK8, RTN3, TUBB3, RPS6, TUBB, RPL23, SH3GL2, TMEM35A, VPS13C, TUBA1B
L6 CT	10	1. GOBP_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE (padj=2.67e-20) 2. GOCC_CYTOSOLIC_RIBOSOME (padj=2.86e-20) 3. GOBP_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ENDOPLASMIC_RETICU (padj=5.51e-20)	—
L6 IT	6	1. GOMF_STRUCTURAL_CONSTITUENT_OF_CYTOSKELETON (padj=3.47e-11) 2. GOMF_GTPASE_ACTIVITY (padj=9.73e-09) 3. GOMF_GUANYL_NUCLEOTIDE_BINDING (padj=2.96e-08)	TUBB3, TUBB, TUBA1B, TUBB2A, TUBB4B, TUBA4A
LAMP5	54	1. GOBP_P_BODY_ASSEMBLY (padj=3.00e-02) 2. GOCC_INTRACELLULAR_PROTEIN_CONTAINING_COMPLEX (padj=3.30e-02)	APPL1, PPP2R2B, WASL, IREB2, COPA, ANKH, GPR107, APPBP2, CNOT7, SPTAN1
LAMP5 LHX6	3	—	—
PAX6	1	—	—
PVALB	89	1. GOBP_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_CONJUGATION_OR_REMOV (padj=1.13e-02) 2. GOBP_PROTEIN_MODIFICATION_BY_SMALL_PROTEIN_CONJUGATION (padj=1.60e-02) 3. GOMF_UBIQUITIN_LIKE_PROTEIN_LIGASE_ACTIVITY (padj=2.84e-02)	PAFAH1B1, APLP2, STIM2, DNM1L, RBL2, PANX1, AKT3, USP10, ATP2A2, SENP1

Figure 2.7: dCoPA reveals reproducible downregulation of modular gene activity in deep-layer intratelencephalic excitatory neurons, PVALB inhibitory neurons, and LAMP5 inhibitory neurons in Alzheimer's disease (AD).

a-b) Dotplots depict the number of bulk coexpression modules identified by dCoPA that were significantly down-regulated in AD vs. CTRL cell types from MTG (**a**) or DFC (**b**) in snRNA-seq data from Gabitto et al.,[29] Liu et al.,[49] or both (gold rows). CTRL / AD modules denote bulk coexpression modules derived from CTRL or AD
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
samples, respectively. **c-d**) Cell type overlap of shared dCoPA genes from (**a**) and (**b**) (i.e., the intersection of all genes comprising the modules in the bottom two rows in each figure panel). **e-f**) Top three results from Gene Ontology[54] enrichment analysis for genes featured in (**c-d**) are shown in third column, and AI-powered multi-agent search results for connections to AD biology are listed in the fourth column.

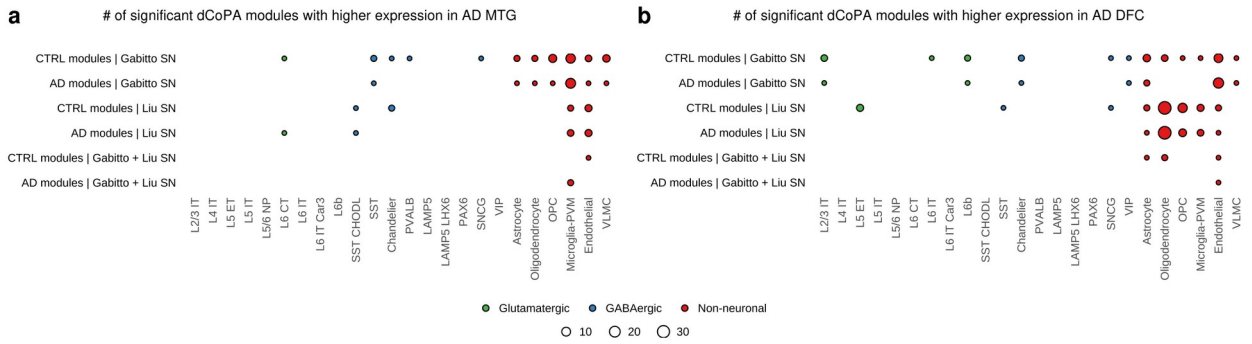


Figure S2.10: Up-regulation of modular gene activity in Alzheimer's disease (AD) occurs mostly in non-neuronal cells.

a-b) Dotplots depict the number of bulk coexpression modules identified by dCoPA that were significantly up-regulated in AD vs. CTRL cell types from MTG (**a**) or DFC (**b**) in snRNA-seq data from Gabitto et al.,[29] Liu et al,[49] or both (gold rows). CTRL / AD modules denote bulk coexpression modules derived from CTRL or AD samples, respectively.

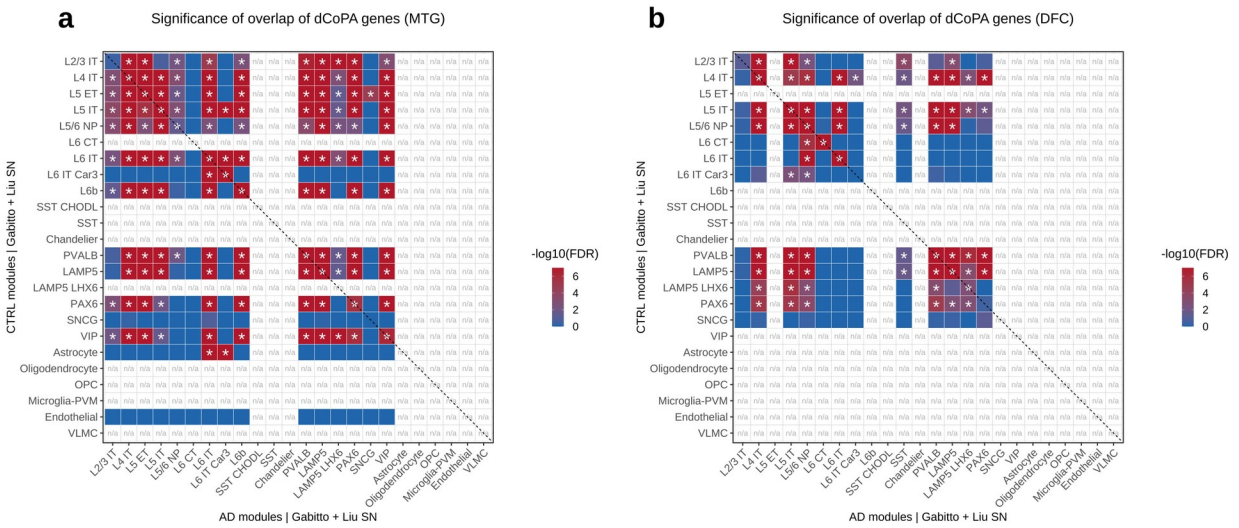
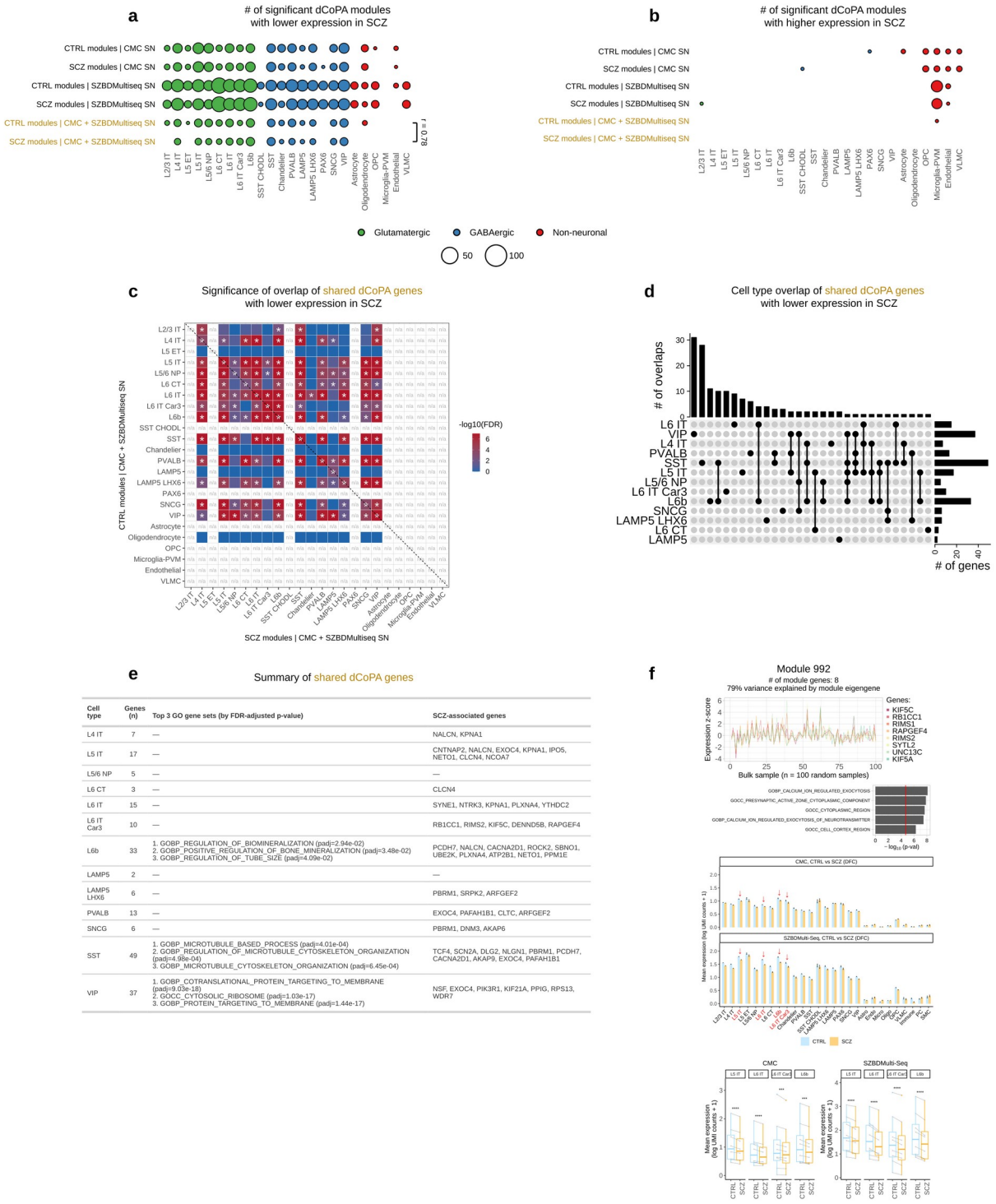


Figure S2.11: Modular gene activity is reproducibly and coordinately down-regulated in specific neuronal subclasses.

a-b) Significance of overlap (one-sided Fisher's exact test) for shared dCoPA genes with lower expression in AD MTG (a) and DFC (b) (i.e., the intersection of all genes for each subclass in the bottom two rows in **Fig. 2.7a** (**a**) and **Fig. 2.7b** (**b**).



Cell type	Genes (n)	Top 3 GO gene sets (by FDR-adjusted p-value)	SCZ-associated genes
L4 IT	7	—	NALCN, KPNA1
L5 IT	17	—	CNTNAP2, NALCN, EXOC4, KPNA1, IPO5, NETO1, CLCN4, NCOA7
L5/6 NP	5	—	—
L6 CT	3	—	CLCN4
L6 IT	15	—	SYNE1, NTRK3, KPNA1, PLXNA4, YTHDC2
L6 IT Car3	10	—	RB1CC1, RIMS2, KIF5C, DENND5B, RAPGEF4
L6b	33	1. GOBP_REGULATION_OF_BIOMINERALIZATION (padj=2.94e-02) 2. GOBP_POSITIVE_REGULATION_OF_BONE_MINERALIZATION (padj=3.48e-02) 3. GOBP_REGULATION_OF_TUBE_SIZE (padj=4.09e-02)	PCDH7, NALCN, CACNA2D1, ROCK2, SBNO1, UBE2K, PLXNA4, ATP2B1, NETO1, PPM1E
LAMP5	2	—	—
LAMP5 LHX6	6	—	PBRM1, SRPK2, ARFGEF2
PVALB	13	—	EXOC4, PAFAH1B1, CLTC, ARFGEF2
SNCG	6	—	PBRM1, DNM3, AKAP6
SST	49	1. GOBP_MICROTUBULE_BASED_PROCESS (padj=4.01e-04) 2. GOBP_REGULATION_OF_MICROTUBULE_CYTOSKELETON_ORGANIZATION (padj=4.98e-04) 3. GOBP_MICROTUBULE_CYTOSKELETON_ORGANIZATION (padj=6.45e-04)	TCF4, SCN2A, DLG2, NLGN1, PBRM1, PCDH7, CACNA2D1, AKAP9, EXOC4, PAFAH1B1
VIP	37	1. GOBP_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE (padj=9.03e-18) 2. GOCC_CYTOSOLIC_RIBOSOME (padj=1.03e-17) 3. GOBP_PROTEIN_TARGETING_TO_MEMBRANE (padj=1.44e-17)	NSF, EXOC4, PIK3R1, KIF21A, PPIG, RPS13, WDR7

Figure 2.8: dCoPA reveals reproducible downregulation of modular gene activity in SST inhibitory neurons, VIP inhibitory neurons, and L6b excitatory neurons in schizophrenia (SCZ).

a-b) Dotplots depict the number of bulk coexpression modules identified by

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)
dCoPA that were significantly down-regulated (**a**) or up-regulated (**b**) in SCZ vs. CTRL cell types from DFC in snRNA-seq data from the SZBDmulti-seq cohort,[65] the CMC cohort,[41] or both (gold rows). CTRL / SCZ modules denote bulk coexpression modules derived from CTRL or SCZ samples, respectively. **c)** Significance of overlap (one-sided Fisher's exact test) for shared dCoPA genes with lower expression in SCZ (i.e., the intersection of all genes for each subclass in the bottom two rows in [**a**]). **d)** Cell type overlap of shared dCoPA genes with lower expression in SCZ (i.e., the same genes analyzed in [**c**]). **e**) Top three results from Gene Ontology[54] enrichment analysis for genes featured in (**c-d**) are shown in third column, and AI-powered multi-agent search results for connections to SCZ biology are listed in the fourth column. **f**) Example of a module exhibiting uniform and reproducible downregulation in deep-layer excitatory neurons in SCZ.

Methods

All analyses were performed in R (v4.4.1) (https://cran.r-project.org/) or Python (v3.9.16).

1. snRNA-seq DE analysis

We downloaded two snRNA-seq datasets[29,30] comprising ~527K neocortical nuclei from dorsolateral prefrontal cortex (DFC) and middle temporal gyrus (MTG) of 12 neurotypical adult human donors. Data were produced using the 10x Chromium V3 (Cv3) platform (Table 2.1).

Table 2.1: Summary of single-nucleus RNA-seq datasets analyzed in this study

Dataset	Platform	# Cases	# Controls	# Nuclei (Cases)	# Nuclei (Controls)	Median # UMIs per Nucleus	Pubmed ID	Data repository	Accession ID
Jorstad et al. 2023	Cv3	0	3	0	347152	16082	37824655	NeMO Archive	nemo:dat-rg2rc5m
Jorstad et al. 2023	SSv4	0	3	0	21858	1710763	37824655	NeMO Archive	nemo:dat-rg2rc5m; nemo:dat-swzf4kc
Gabitto et al. 2024	Cv3	75	9	2224098	321832	18921	39402379	Synapse (AD Knowledge Portal)	syn26223298
Liu et al. 2025	Cv3	51	55	527907	626175	5723	40752494	MIT (compbio.mit.edu)	compbio.mit.edu
Emani et al. 2024 (CMC)	Cv3	47	53	214234	287783	2846	38781369	Synapse (PsychENCODE)	syn51111084
Emani et al. 2024 (SZBDMulti-Seq)	Cv3	24	24	171313	173570	9997	38781369	Synapse (PsychENCODE)	syn51111084
Morabito et al. 2021	Cv3	11	7	38676	22796	6387	34239132	GEO	GSE174367

1.1 Construction of donor- and subclass-level pseudobulk samples from snRNA-seq data

To perform differential expression (DE) analysis, we used a pseudobulk approach to avoid pseudoreplication bias and excessive false discoveries that can result from performing DE directly on individual cells or nuclei from the same donors.[16-18] Raw UMI counts were summed for each subclass (n = 24) from each donor (n = 12) in each brain

region, resulting in 72 DFC or MTG pseudobulk samples from Jorstad et al.[30] and 216 DFC or MTG pseudobulk samples from Gabitto et al.[29]

1.2 Differential expression (one subclass vs. all other subclasses)

DE was performed using the R packages edgeR[33] (v4.4.2) and DESeq2[34] (1.46.0). For edgeR we used the glmLRT (Generalized Linear Model Likelihood Ratio Test) pipeline without filterByExpr and with robust = TRUE for the estimateDisp function. For DESeq2, test = "LRT" was utilized with the reduced model defined using donors only. DE was performed after filtering genes to those present in both Jorstad et al.[30] and Gabitto et al.[29] (n = 20,568 for DFC and n = 20,892 for MTG). For each subclass, marker genes were defined as those that were significantly upregulated (FDR < .05 and $log_2(fold\ \ change) > 0$) and unique (i.e., significantly upregulated in only one subclass).

2. Pseudobulk modeling

In addition to the datasets used for DE analysis, we downloaded additional snRNA-seq data from neurotypical adult human neocortex sample cohorts from Jorstad et al.,[30] including Cv3 primary visual cortex (V1) (n = three donors and 141,836 nuclei), SMART-seq V4 (SSv4) DFC (n = three donors and 4,551 nuclei), and SSv4 MTG (n = three donors and 17,307 nuclei).

2.1 Pseudobulk dataset creation for modeling gene expression vs. cell type abundance

For each snRNA-seq dataset cohort, pseudobulk samples were created by randomly sampling and summing 10% of all nuclei to create synthetic samples. We matched the number of samples (n = 1,518) and genes (n = 18,913) in each pseudobulk dataset to match the integrated bulk control dataset (**Fig. 2.3a**). The cell type

composition of each pseudobulk sample was recorded to create cell type abundance vectors for linear modeling of gene expression.

2.2 Linear modeling of pseudobulked gene expression by cell type abundance

To model pseudobulked gene expression as a function of cell type abundance, we performed multiple linear regression with each gene (γ_s) in each pseudobulk dataset consisting of s samples as response and cell type abundance vectors (a) as predictors using the lm function in R:

$$\gamma_s = \beta_0 + \sum_{c=1}^{C} \beta_c a_{cs} + \varepsilon_s \quad (1)$$

where a_{cs} represents the relative abundance of cell type c in sample s, β_c represents the regression coefficient for cell type c, β_0 represents the model intercept, and ε_s represents the error term. Cell types (c) were defined using either subclass (n = 24) or supertype (n = 136) definitions. Modeling performance was quantified with Adjusted R^2 or the root mean squared error (RMSE):

$$RMSE = \sqrt{1/n \sum_{s=1}^{n} (\gamma_s - \hat{\gamma}_s)^2} \quad (2)$$

Where γ_s represents real gene expression in sample s, $\hat{\gamma}_s$ represents predicted gene expression in sample s, and *n* represents the total number of samples.

3. CoPA

CoPA (**Co**variation **P**rojection **A**nalysis) is a pipeline for integrating genome-wide covariation analysis of bulk tissue samples with analysis of pseudobulked cell types from single-cell or single-nucleus datasets. The full pipeline starts with a large bulk dataset as input, followed by genome-wide covariation analysis, followed by covariation

module projection onto pseudobulked cell types from single-cell or single-nucleus datasets (**Fig. 2.3a**).

3.1 Bulk RNA-seq pre-processing

We downloaded bulk RNA-seq data from six studies[38-44] representing adult (≥18 years) human neurotypical DFC samples. After filtering and quality control (see below), we obtained a total of 1,518 samples. All pre-processing was performed from FASTQ files with the exception of GTEx (see below). Raw sequence reads were downloaded from Synapse (https://www.synapse.org/) using the following identifiers: syn8612097 (Religious Orders Study and Memory and Aging Project Study, or ROSMAP), syn18134196 (Common Mind Consortium, or CMC, split between the MSSM/Penn/Pitt cohort and the NIMH Human Brain Collection Core, or HBCC, cohort), syn8227833 (BrainSeq Phase 1), and syn7062404 (BrainGVEX). Data from GTEx (v8) was downloaded from the AnVIL repository (https://anvil.terra.bio). Data from the North American Brain Expression Consortium (NABEC) was downloaded from dbGaP (https://dbgap.ncbi.nlm.nih.gov/) with the accession ID phs001353.v1.p1 (Table 2.2).

Table 2.2: Bulk RNA-seq datasets analyzed in this study

Dataset	First author last name	Last author last name	Journal	Year	PMID	Data Repository	Accession ID or URL	Platform	No of samples filtered
ROSMAP (CTRL)	Mostafavi	De Jager	Nature Neuroscience	2018	29802388	Synapse (AD Knowledge Portal)	syn3219045	RNAseq	347
ROSMAP (AD)	Mostafavi	De Jager	Nature Neuroscience	2018	29802388	Synapse (AD Knowledge Portal)	syn3219045	RNAseq	248
CMC	Hoffman	Roussos	Scientific Data	2019	31551426	Synapse (CommonMind)	syn2759792	RNAseq	285
CMC_HBCC	Hoffman	Roussos	Scientific Data	2019	31551426	Synapse (CommonMind)	syn2759792	RNAseq	162
BrainGVEX	Gandal	Geschwind	Science	2018	30545856	Synapse (PsychENCODE)	syn3270015	RNAseq	275
Brainseq (CTRL)	Jaffe	Weinberger	Nature Neuroscience	2018	30050107	Synapse (PsychENCODE)	syn12299750	RNAseq	190

Dataset	First author last name	Last author last name	Journal	Year	PMID	Data Repository	Accession ID or URL	Platform	No of samples filtered
Brainseq (SCZ)	Jaffe	Weinberger	Nature Neuroscience	2018	30050107	Synapse (PsychENCODE)	syn12299750	RNAseq	171
GTEx	GTEx Consortium	GTEx Consortium	Science	2020	32913098	AnVIL (dbGaP)	phs000424.v8.p2	RNAseq	190
NABEC	Dillman	Cookson	Scientific Reports	2017	29203886	dbGaP	phs001353.v1.p1	RNAseq	69

Reads were quantified from raw expression files using kallisto[57] (v0.44.0) with default settings for paired-end data. A kallisto index was built using the GENCODE v39 transcriptome (GRCh38).[58] BAM files from GTEx were converted to FASTQ files using the run_SamToFastq.py script from the Broad Institute's gtex-pipeline repository (https://github.com/broadinstitute/gtex-pipeline). Transcript-level abundances were summarized to gene-level counts with tximport[59] (v1.34.0) using a GENCODE-derived (v39) transcript-to-gene mapping.

Each bulk dataset was processed using the SampleNetwork[46] R function, a bulk expression pre-processing tool designed for sample outlier removal and batch correction. Outliers were defined as samples with connectivity more than four standard deviations below the mean connectivity of all samples over all genes ($Z.K <$ -4). Outliers were iteratively removed until no samples with $Z.K <$ -4 remained. Next, samples were batch corrected for library batch if the corresponding information was available. Batch correction was performed using the Combat[45] R function, which is built into SampleNetwork. Finally, genes with zero variance were removed.

After outlier removal and batch correction, all datasets were concatenated into a single dataset after aligning to shared genes (n = 18,913). SampleNetwork was then applied again to the integrated dataset to remove outliers and to batch correct by

individual dataset. The final batch-corrected, integrated bulk control dataset was used as input for downstream analyses.

In addition to the integrated bulk control dataset, two additional disease-specific bulk datasets were prepared and utilized for downstream analyses: 248 samples from ROSMAP[53] with Alzheimer's disease (AD) and 171 samples from Brainseq Phase 1[42] with schizophrenia (SCZ). These datasets were quantified from FASTQ files using kallisto and processed by SampleNetwork in an identical fashion to each individual dataset of the integrated bulk control dataset.

3.2 Unsupervised co-expression module detection with iterative sequestering

We used a variant of the coexpression module detection approach described by Kelley et al.[22] We implemented a 'greedy' approach that iteratively finds the most highly correlated groups of genes and removes them prior to the next iteration, thereby increasing the discoverability of modules that might be masked by more highly correlated groups of genes. Each iteration of the algorithm proceeded as follows: first, pairwise Pearson correlations were calculated for all genes over all samples of the integrated bulk control dataset (n = 1,518) to produce a genome-wide similarity matrix. Next, the similarity matrix was hierarchically clustered with complete linkage using the flashClust[60] package (v1.1.2) using $1 - cor$ as the distance measure. The resulting dendrogram was then cut at a height corresponding to the top 0.01% of pairwise correlations to produce an initial set of modules, which were then filtered for modules with a minimum size of 10 genes - this threshold was enforced for all subsequent clustering steps of the algorithm. After recording this initial set of modules, the seed genes representing these modules were removed from the similarity matrix, thus

concluding the iteration. The subsequent iteration would then proceed on the trimmed similarity matrix.

For a given dendrogram cut height, the number of modules found would naturally decrease with each iteration until no modules were found. At this point, the algorithm would relax the cut height parameter and then repeat the entire iterative loop again. Initial cut height was set at the top 0.01% of all pairwise correlations, then was lowered in a stepwise fashion according to the following thresholds: 0.1%, 1%; then increasing by 1% until 10%; then increasing by 5% each iteration until a statistically determined lower threshold for minimum module seed gene pairwise correlation was reached. This threshold was defined as the 99th percentile of the distribution of minimum pairwise correlations calculated from randomly generated modules of size 10 (n = 10,000). Once the cut height exceeded this threshold, the algorithm would stop, concluding the module discovery process.

Expanded module definitions ("topmodposbc") were then calculated for all modules. First, each module was summarized by its module eigengene,[47] defined as the first principal component of the gene expression vectors defining the module. Genes were then ranked for each module by the WGCNA measure of intramodular connectivity (k_{ME}, defined as the Pearson correlation of a gene with a module eigengene[47]), and kept if the k_{ME} value was significant according to a Bonferroni-corrected threshold (where the number of comparisons is defined by the number of genes times the number of modules). Finally, genes were kept for a given module if the gene had the highest k_{ME} value for that module compared to all others. Defining modules this way resulted in more genes associated with each module on average, but

a fraction of the modules had fewer topmodposbc genes than the 10 gene threshold enforced for module seed genes (**Fig. 2.4q**).

Before continuing with downstream analyses, modules underwent additional quality control filtering. First, modules with three or fewer genes according to the topmodposbc expanded definition were discarded. This resulted in 135 modules discarded from the integrated bulk control modules, 117 modules discarded from the ROSMAP AD modules, and 64 modules discarded from the Brainseq SCZ modules. Second, for the integrated bulk control modules specifically, the significance of pairwise correlations of topmodposbc genes in each input dataset for the integrated bulk control dataset was calculated for all modules by comparing real pairwise correlations of a given module in a given dataset to the pairwise correlations of genes from randomly generated modules (n = 1,000), matched to the size of the module. Modules that did not have pairwise correlations significantly greater than chance in at least two datasets were discarded. This resulted in seven additional modules being discarded from the integrated bulk control modules. The final module count for each dataset was as follows: 1,016 modules for the integrated bulk control dataset, 1,010 modules for the ROSMAP AD dataset, and 1,083 modules for the Brainseq SCZ dataset.

3.3 Gene set enrichment analysis

To clarify module identity and function, we performed enrichment analyses of module topmodposbc genes with two collections of gene sets: the Molecular Signatures Database (MSigDB v7.4)[48] from the Broad Institute (https://www.gsea-msigdb.org/gsea/msigdb) and a curated list of cell type gene sets sourced from a wide range of bulk RNA expression and SC/SN RNAseq datasets

available in OMICON (https://theomicon.ucsf.edu). We used the following human collections from the MSigDB database: H (hallmark gene sets), C2 (curated gene sets), C5 (ontology gene sets) and C8 (cell type signature gene sets). Enrichment tests were performed with a one-sided Fisher's exact test using the base R phyper function.

3.4 Projecting modules onto snRNA-seq data

We projected co-expression modules onto a series of snRNA-seq datasets. In addition to the previously mentioned Jorstad et al.[30] and Gabitto et al.[29] datasets, we downloaded two additional datasets from studies of adult human neurotypical DFC. Data from Liu et al. 2025[49] were downloaded from https://compbio.mit.edu/AD_Multiomic_MultiRegion/, consisting of 626,175 nuclei from 55 individuals. This resource aggregates nuclei generated by Liu et al. with previously published data from Mathys et al.,[61] Xiong et al.,[62] and Mathys et al.[63] Data from Morabito et al. 2021[31] were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) using the accession GSE174367, consisting of 22,796 nuclei from seven individuals.

To project modules onto pseudobulked neocortical subclasses from snRNA-seq datasets, snRNA-seq data were log-transformed + 1 (log1p). Then, for each subclass and each module (defined using either seed genes or topmodposbc genes), mean expression was calculated over all nuclei from each subclass (**Fig. 2.3e**). To calculate the relative expression index (REI), these projection indices were then normalized by the mean expression of all genes in each subclass, then scaled so that the maximum REI value was 1 (**Fig. 2.3f**).

3.5 Cross-dataset subclass mapping

To enable subclass-specific cross-dataset comparisons between Gabitto et al.[29] and Liu et al.,[49] we mapped all nuclei from Liu et al. to Gabitto et al. subclasses using the following strategy: first, raw UMI counts for both datasets were log-transformed +1 (log1p). Then, metacells for each Gabitto et al. subclass (n = 24) were created by averaging expression vectors for all nuclei belonging to a given subclass to produce a gene by metacell matrix. Then, the Pearson correlations between each nucleus in Liu et al. and each metacell from Gabitto et al. were calculated, and subclass labels were mapped to each nucleus based on the most highly correlated metacell (**Fig. S2.6**). We repeated this process for DFC and MTG regions separately.

3.6 Consensus meta-module clustering

To identify patterns among module projections, we hierarchically clustered module projections to produce “meta-modules” that summarize common archetypes of gene activity across neocortical subclasses. Before doing so, to ensure that we identified projection patterns that were reproducible across datasets, we calculated a consensus minimum similarity matrix from REI projections of the integrated bulk control modules onto Jorstad et al.[30] and control samples from Gabitto et al.[29] (**Fig. 2.5a**). Specifically, for each dataset we calculated module-by-module similarity matrices by calculating pairwise Pearson correlations of all modules over all REI projection indices. The resulting two matrices were then collapsed into a single consensus matrix by calculating the parallel minimum across both matrices using the pmin R function, which in essence takes the minimum correlation value at each position after overlaying both matrices on top of one another. This consensus approach ensures that any co-expression relationships are reproducible across both input datasets.

This consensus similarity matrix was then processed using previously described methods. We clustered the consensus similarity matrix using the flashclust (v 1.1.2) implementation of complete linkage hierarchical clustering with $1-cor$ as a distance measure. The resulting dendrogram was then cut at a height corresponding to the top 5% of all pairwise correlations, producing a series of meta-modules where each meta-module represents a group of highly correlated module projection patterns. Meta-modules were kept if the minimum size of the meta-module was ≥ 3. Furthermore, meta-modules were summarized by their "eigenmodules", defined as the first principal component of the module REI projections corresponding to each meta-module ("seed modules"). This is conceptually identical to summarizing gene co-expression modules by their eigengenes. Highly similar meta-modules were merged if their correlations exceeded 0.95 by combining the seed modules of each meta-module into a single meta-module. This pipeline was repeated for two regions: DFC (**Fig. 2.5b-d**) and MTG (**Fig. S2.8**).

In parallel, subclasses were clustered by applying a similar clustering strategy to the transpose of the previous analysis: subclass-by-subclass similarity matrices were produced by calculating pairwise Pearson correlations of all subclass REI projection indices over all modules. Like before, we used pmin to produce a consensus minimum subclass-by-subclass similarity matrix. This matrix was used as input into complete linkage hierarchical clustering via the flashclust package.

To visualize meta-modules, heatmaps of eigenmodules calculated from projection data from either Jorstad et al.[30] or Gabitto et al.[29] were displayed with the consensus meta-module dendrogram overlaid on top of each heatmap (**Fig. 2.5b-d,**

Fig. S2.8). As eigenmodules are calculated using principal component analysis (PCA), the values of eigenmodules are arbitrary units that range from -1 to 1. Subclasses (y-axis of the heatmaps) were clustered according to the consensus subclass dendrogram. For each node of the meta-module dendrogram above a certain arbitrary threshold (distance = 0.3), nodes were labeled with two summary statistics: the percentage of all modules represented by all leaves under a given node ("m"), and the percentage of all module genes represented by those modules ("g"). The same statistics for each individual leaf are displayed as barplots below the dendrograms.

4. dCoPA

In addition to the neurotypical snRNA-seq samples from Gabitto et al.[29] and Liu et al.,[49] we ran CoPA on disease samples from the same datasets. Gabitto et al. consists of 75 donors with Alzheimer's disease representing 2,224,098 nuclei, while Liu et al. consists of 51 donors with Alzheimer's disease representing 527,907 nuclei. We also downloaded data from the brainSCOPE resource[55] (https://brainscope.gersteinlab.org/), which profiles adult human DFC samples from the PsychENCODE Consortium[64] (https://www.psychencode.org/). Of the studies aggregated by the brainSCOPE resource, we analyzed studies from the CommonMind Consortium[41] (53 control donors representing 287,783 nuclei and 47 schizophrenia donors representing 214,234 nuclei) and SZBDMulti-Seq[65] (24 control donors representing 173,570 nuclei and 24 schizophrenia donors representing 171,313 nuclei).

4.1 dCoPA overview

We leveraged CoPA projections to find disease-specific gene expression dysregulation at module-level resolution. To achieve this, we designed an

accompanying pipeline (called “dCoPA” or differential CoPA). dCoPA selects a group of candidate disease-related modules based on the application of a series of filtering criteria to CoPA projections that ensure that any discovered disease associations are significant, interpretable, and reproducible (**Fig. 2.6a**).

The first filtering step was to determine whether module projections were significantly different between case and control samples for a given subclass. To do this, we projected modules from the integrated bulk control dataset onto case and control samples in mean log space. For each module (n = 1,016) and subclass (n = 24), we calculated the Euclidean distance between the mean expression vectors of all module genes in cases and controls. We compared this distance to a distribution of randomly generated distances, created by randomly generating and projecting 10,000 modules for each real module (for a total of 10,000 times 1,016 modules) that were matched in size to the corresponding real module. Each distance was assigned an empirical p-value representing the proportion of random distances smaller than the real distance. P-values were then adjusted according to the Benjamini-Hochberg procedure.

In addition to significance, we also filtered modules according to the directionality of module genes. We selected significant modules where module topmodposbc genes were either uniformly up- or down-regulated in disease samples relative to control samples. Any module where directionality was not consistent (e.g., some module genes showing higher expression in disease samples with others showing lower expression) was filtered out. Our rationale for applying this additional filtering step was to maximize the interpretability of the final list of dCoPA-selected modules.

4.2 Analysis of the rate of discovery of significant projections from random modules

To ensure that the significance of projections could not be spuriously found in the data, we generated 1,000 random module networks, each consisting of the same number of modules (n = 1,016) as the real module network generated from the integrated bulk control dataset. Random modules were generated by randomly sampling all genes (n = 18,913), where each module of each random network was matched in size to the corresponding module in the real network. For each random network, we calculated the percentage of modules marked significant by dCoPA (boxplots, **Fig. S2.9a-d**), and we calculated the average percentage of subclasses per module marked significant by dCoPA (boxplots, **Fig. S2.9e-h**). We calculated the same percentages for the real network (red dots, **Fig. S2.9**).

4.3 Comparison of control- versus disease-derived dCoPA modules

To assess the reproducibility of dCoPA modules, we applied dCoPA to projections produced from two sets of modules (derived from either the integrated bulk control dataset or the AD ROSMAP dataset[53]) and two snRNA-seq datasets (Gabitto et al.[29] and Liu et al.[49]) split by region (DFC and MTG), for a total of eight sets of candidate dCoPA modules (**Fig. 2.7a-b**).

To find which dCoPA modules were common to both datasets, we identified the intersection per subclass between dCoPA modules found from Gabitto et al. versus dCoPA modules found from Liu et al. (bottom two rows of **Fig. 2.7a-b**). Subclasses were matched between the two datasets based on the mapping strategy described previously. We did this separately for each set of modules and each region.

To compare dCoPA modules derived from the integrated bulk control dataset ("control-derived") to those derived from the AD ROSMAP dataset ("AD-derived"), we first found common dCoPA modules between datasets as described above. We then summarized common dCoPA modules for each subclass by the union of their topmodposbc genes. Finally, we performed one-sided Fisher's exact tests comparing control-derived genes to AD-derived genes for each subclass (**Fig. S2.11**).

To assess the overlap of dCoPA genes across subclasses, we first found common genes that overlapped between control-derived and AD-derived genes per subclass. To visualize the overlap of these genes by subclass, we used these genes as input for the UpSet function from the ComplexHeatmap[66] package (v2.22.0). Combinations where one or fewer genes overlapped between subclasses were not plotted for readability.

We repeated all of the above analyses for SCZ-derived modules and projections. We applied dCoPA to projections produced from two sets of modules (derived from either the integrated bulk control dataset or the SCZ Brainseq Phase 1 dataset[42]) and two snRNA-seq datasets (CMC[41] and SZBDMulti-Seq[65] from the brainSCOPE resource[55]).

4.4 Summary of dCoPA genes

We created summary tables for dCoPA genes by subclass. For each subclass, we performed enrichment analysis for dCoPA genes associated with that subclass using Gene Ontology gene sets obtained from the Molecular Signatures Database (MSigDB v7.4). Enrichment tests were performed by using a one-sided Fisher's exact test using

the base R phyper function. The top three gene sets by Benjamini-Hochberg adjusted p-values per subclass were shown.

dCoPA genes were annotated for association with Alzheimer's disease and schizophrenia with a single pipeline (Claude Sonnet 5, run via Claude Code [https://www.anthropic.com/claude-code]), applied independently to each disease. The reproducible cross-dataset dCoPA gene lists (i.e., genes belonging to modules that were significantly and uniformly dysregulated in the same cell type[s]) were associated with disease via one of three tracks. First, genes present in a curated disease database (OMIM,[67] Open Targets Platform,[68] ClinVar,[69] and, for AD, AlzForum [https://www.alzforum.org]) were admitted directly. Second, for all remaining genes, NCBI E-utilities (https://www.ncbi.nlm.nih.gov/books/NBK25497/) deterministically retrieved the total disease hit count and the top five most relevant PubMed abstracts for "GENE"[TIAB] AND "<disease>"[TIAB] (2000–present); because abstracts and PMIDs were fetched directly from NCBI, citation hallucination was impossible by construction. Claude Sonnet 5 then classified a gene as disease-linked only if a retrieved abstract established a direct mechanistic link to one of the disease's mechanism categories (13 for AD, 14 for SCZ), rejecting co-mentions, differential-expression-only findings, and gene-symbol homonyms, and citing only fetched PMIDs. Genes missed by the exact symbol were re-queried using NCBI gene aliases under a homonym-specificity guard. Third, three AD genes whose literature refers to a shared protein-family name (PPP3CB/calcineurin, SPTAN1/spectrin, and EDEM3) were preserved from prior manual curation; schizophrenia required none. Each gene was annotated with its total PubMed hit count and, for AD, its Open Targets association score (disease

MONDO_0004975), then intersected with the significant and reproducible dCoPA gene set and annotated with its normal and disease module membership, associated neuronal subclass(es), and direction of module change. The final results comprised 74 (DFC) and 215 (MTG) genes for AD and 51 (DFC) genes for schizophrenia.

4.5 CoPA Cabana (R Shiny app)

We made CoPA and dCoPA results explorable by developing an interactive web application in R with the Shiny framework (bslib UI, plotly interactive graphics), which was deployed on shinyapps.io (oldhamlab.shinyapps.io/copacabana). The application comprises four tabs. The CoPA tab enables browsing and gene-based searching for gene coexpression modules identified in our integrated control cohort (**Fig. 2.3**), displaying the same information portrayed in **Fig. 2.3b-f** for all modules. The dCoPA tab visualizes all dCoPA-implicated modules for all disease comparisons explored in **Figs. 2.7**-**8.** The OMICON tab allows users to project coexpression modules downloaded from OMICON (https://theomicon.ucsf.edu) onto a number of select snRNA-seq reference atlases. The Gene Projection tab reports per-gene cell-type expression across reference datasets.

Competing interests

The authors declare no competing interests.

Data availability

All analyses are based on published datasets accessed as described in Methods. Results of our analyses are available in supplementary tables and on the CoPA Cabana web site that accompanies this study (https://oldhamlab.shinyapps.io/copacabana/).

Accession identifiers and repositories for all datasets are provided in Table S1 (single-nucleus RNA-seq) and Table S8 (bulk RNA-seq).

Code availability

Code to reproduce our analyses is available in the Oldham lab GitHub repository (https://github.com/oldham-lab/kang-oldham-2026).

References

1. Bahar Halpern, K. *et al.* Nuclear Retention of mRNA in Mammalian Tissues. *Cell Rep* **13**, 2653-62 (2015).

2. Battich, N., Stoeger, T. & Pelkmans, L. Control of Transcript Variability in Single Mammalian Cells. *Cell* **163**, 1596-610 (2015).

3. Gorin, G. & Pachter, L. Length biases in single-cell RNA sequencing of pre-mRNA. *Biophys Rep (N Y)* **3**, 100097 (2023).

4. Tang, W., Jorgensen, A.C.S., Marguerat, S., Thomas, P. & Shahrezaei, V. Modelling capture efficiency of single-cell RNA-sequencing data improves inference of transcriptome-wide burst kinetics. *Bioinformatics* **39**(2023).

5. Yang, A.C. *et al.* A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. *Nature* **603**, 885-892 (2022).

6. Bloom, J.D. Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ* **6**, e5578 (2018).

7. DePasquale, E.A.K. *et al.* DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Rep* **29**, 1718-1727 e8 (2019).

8. Schriever, H. & Kostka, D. Vaeda computationally annotates doublets in single-cell RNA sequencing data. *Bioinformatics* **39**(2023).

9. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291 e9 (2019).

10. Fleming, S.J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat Methods* **20**, 1323-1335 (2023).

11. Schaefer, N.K., Pavlovic, B.J., Schmitz, M.T. & Pollen, A.A. CellBouncer, a unified toolkit for single-cell demultiplexing and ambient RNA analysis, reveals hominid mitochondrial incompatibilities. *Cell Genom* **6**, 101275 (2026).

12. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol* **21**, 57 (2020).

13. Young, M.D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**(2020).

14. Luo, Q. & Zhang, H. Emergence of Bias During the Synthesis and Amplification of cDNA for scRNA-seq. *Adv Exp Med Biol* **1068**, 149-158 (2018).

15. Rich, J.M. *et al.* The impact of package selection and versioning on single-cell RNA-seq analysis. *Cell Syst* **17**, 101560 (2026).

16. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* **24**, 550-572 (2023).

17. Junttila, S., Smolander, J. & Elo, L.L. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data. *Brief Bioinform* **23**(2022).

18. Squair, J.W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* **12**, 5692 (2021).

19. Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nat Biotechnol* **39**, 1008-1016 (2021).

20. Zhang, M.J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat Commun* **11**, 774 (2020).

21. Hawrylycz, M. *et al.* Canonical genetic signatures of the adult human brain. *Nat Neurosci* **18**, 1832-44 (2015).

22. Kelley, K.W., Nakao-Inoue, H., Molofsky, A.V. & Oldham, M.C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat Neurosci* **21**, 1171-1184 (2018).

23. Lui, J.H. *et al.* Radial glia require PDGFD–PDGFRB signalling in human but not mouse neocortex. *Nature* **515**, 264-268 (2014).

24. Mistry, M., Gillis, J. & Pavlidis, P. Meta-analysis of gene coexpression networks in the post-mortem prefrontal cortex of patients with schizophrenia and unaffected controls. *BMC Neurosci* **14**, 105 (2013).

25. Oldham, M.C. *et al.* Functional organization of the transcriptome in human brain. *Nat Neurosci* **11**, 1271-1282 (2008).

26. Ponomarev, I., Wang, S., Zhang, L., Harris, R.A. & Mayfield, R.D. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. *J Neurosci* **32**, 1884-97 (2012).

27. Raju, C.S. *et al.* Secretagogin is Expressed by Developing Neocortical GABAergic Neurons in Humans but not Mice and Increases Neurite Arbor Size and Complexity. *Cereb Cortex* **28**, 1946-1958 (2018).

28. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380-4 (2011).

29. Gabitto, M.I. *et al.* Integrated multimodal cell atlas of Alzheimer's disease. *Nat Neurosci* **27**, 2366-2383 (2024).

30. Jorstad, N.L. *et al.* Transcriptomic cytoarchitecture reveals principles of human neocortex organization. *Science* **382**, eadf6812 (2023).

31. Morabito, S. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet* **53**, 1143-1155 (2021).

32. Network, B.I.C.C. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86-102 (2021).

33. Chen, Y., Chen, L., Lun, A.T.L., Baldoni, P.L. & Smyth, G.K. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Res* **53**(2025).

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

35. French, L., Suresh, H. & Gillis, J. Cluster replicability in single-cell and single-nucleus atlases of the mouse brain. *Nat Commun* (2026).

36. Nakatsuka, N. *et al.* Improving reproducibility of differentially expressed genes in single-cell transcriptomic studies of neurodegenerative diseases through meta-analysis. *Nat Commun* **16**, 7436 (2025).

37. Nguyen, H.C.T., Baik, B., Yoon, S., Park, T. & Nam, D. Benchmarking integration of single-cell differential expression. *Nat Commun* **14**, 1570 (2023).

38. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

39. Dillman, A.A. *et al.* Transcriptomic profiling of the human brain reveals that altered synaptic gene expression is associated with chronological aging. *Sci Rep* **7**, 16890 (2017).

40. Gandal, M.J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**(2018).

41. Hoffman, G.E. *et al.* CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci Data* **6**, 180 (2019).

42. Jaffe, A.E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**, 1117-1125 (2018).

43. Mostafavi, S. *et al.* A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat Neurosci* **21**, 811-819 (2018).

44. BrainSeq: A Human Brain Genomics Consortium. BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders. *Neuron* **88**, 1078-1083 (2015).

45. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).

46. Oldham, M.C., Langfelder, P. & Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst Biol* **6**, 63 (2012).

47. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol* **4**, e1000117 (2008).

48. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425 (2015).

49. Liu, Z. *et al.* Single-cell multiregion epigenomic rewiring in Alzheimer's disease progression and cognitive resilience. *Cell* **188**, 4980-5002 e29 (2025).

50. Vialle, R.A. *et al.* Structural variants linked to Alzheimer's disease and other common age-related clinical and neuropathologic traits. *Genome Med* **17**, 20 (2025).

51. Kumar, A. *et al.* Genetic correlation analysis identifies TMEM106B, ACE, and ERC2 as genetic loci shared between Alzheimer's disease and primary psychiatric disorders. *Alzheimers Dement* **22**, e71278 (2026).

52. Cholerton, B. *et al.* Genome wide association study meta-analysis of neuropathologic lesions of Alzheimer's disease and related dementias in a multi-site autopsy cohort. *PLoS Genet* **22**, e1012170 (2026).

53. Bennett, D.A., Schneider, J.A., Arvanitakis, Z. & Wilson, R.S. Overview and findings from the religious orders study. *Curr Alzheimer Res* **9**, 628-45 (2012).

54. Gene Ontology, C. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Res* **54**, D1779-D1792 (2026).

55. Emani, P.S. *et al.* Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**, eadi5199 (2024).

56. Kampmann, M. Molecular and cellular mechanisms of selective vulnerability in neurodegenerative diseases. *Nat Rev Neurosci* **25**, 351-371 (2024).

57. Bray, N.L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**, 525-7 (2016).

58. Mudge, J.M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res* **53**, D966-D975 (2025).

59. Soneson, C., Love, M.I. & Robinson, M.D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015).

60. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

61. Mathys, H. *et al.* Single-cell multiregion dissection of Alzheimer's disease. *Nature* **632**, 858-868 (2024).

62. Xiong, X. *et al.* Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422-4437 e21 (2023).

63. 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).

64. Psych, E.C. *et al.* The PsychENCODE project. *Nat Neurosci* **18**, 1707-12 (2015).

65. Ruzicka, W.B. *et al.* Single-cell multi-cohort dissection of the schizophrenia transcriptome. *Science* **384**, eadg5136 (2024).

66. Gu, Z. Complex heatmap visualization. *Imeta* **1**, e43 (2022).

67. Amberger, J.S., Bocchini, C.A., Scott, A.F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res* **47**, D1038-D1043 (2019).

68. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res* **51**, D1353-D1359 (2023).

69. Landrum, M.J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res* **48**, D835-D844 (2020).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Eugene Kang
7DF188DD07E542C...
Author Signature

08/05/2026
Date