Title

An integrated single-nucleus and spatial transcriptomics atlas reveals the molecular landscape of the human hippocampus

Authors

Erik D. Nelson[1,2], Madhavi Tippani[1], Anthony D. Ramnauth[1,3], Heena R. Divecha[1], Ryan A. Miller[1], Nicholas J. Eagles[1], Elizabeth A. Pattie[1], Sang Ho Kwon[1,3,4], Svitlana V. Bach[1], Uma M. Kaipa[1], Jianing Yao[5], Joel E. Kleinman[1,6], Leonardo Collado-Torres[1,5], Shizhong Han[1,6,8], Kristen R. Maynard[1,3,6], Thomas M. Hyde[1,6,7], Keri Martinowich[1,3,6,8,9], Stephanie C. Page[1,6+], Stephanie C. Hicks[5,10,11,12,+]

Affiliations

1. Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA
2. Cellular and Molecular Medicine Graduate Program, Johns Hopkins School of Medicine, Baltimore, MD, USA
3. Solomon H. Snyder Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, USA
4. Biochemistry, Cellular, and Molecular Biology Graduate Program, Johns Hopkins School of Medicine, Baltimore, MD, USA
5. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
6. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA
7. Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD, USA
8. Department of Genetic Medicine, Johns Hopkins School of Medicine, MD, USA
9. Johns Hopkins Kavli Neuroscience Discovery Institute, Baltimore, MD, USA
10. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
11. Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
12. Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

+co-corresponding authors

Correspondence

Stephanie Cerceo Page
stephanie.page@libd.org

Stephanie C. Hicks
shicks19@jhu.edu

Abstract

The hippocampus contains many unique cell types, which serve the structure's specialized functions, including learning, memory and cognition. These cells have distinct spatial topography, morphology, physiology, and connectivity, highlighting the need for transcriptome-wide profiling strategies that retain cytoarchitectural organization. Here, we generated spatially-resolved transcriptomics (SRT) and single-nucleus RNA-sequencing (snRNA-seq) data from adjacent tissue sections of the anterior human hippocampus across ten adult neurotypical donors. We defined molecular profiles for hippocampal cell types and spatial domains. Using non-negative matrix factorization and transfer learning, we integrated these data to define gene expression patterns within the snRNA-seq data and infer the expression of these patterns in the SRT data. With this approach, we leveraged existing rodent datasets that feature information on circuit connectivity and neural activity induction to make predictions about axonal projection targets and likelihood of ensemble recruitment in spatially-defined cellular populations of the human hippocampus. Finally, we integrated genome-wide association studies with transcriptomic data to identify enrichment of genetic components for neurodevelopmental, neuropsychiatric, and neurodegenerative disorders across cell types, spatial domains, and gene expression patterns of the human hippocampus. To make this comprehensive molecular atlas accessible to the scientific community, both raw and processed data are freely available, including through interactive web applications.

Keywords
hippocampus, spatially-resolved transcriptomics, postmortem human brain, single-nucleus RNA-sequencing

1 | Introduction

Spatial organization in brain tissue reflects differences in cellular composition, neural connectivity, and function. Recent advancements [1–3] in spatially-resolved transcriptomics (SRT) have revealed how unique patterns of gene expression map onto topography of different brain regions. These techniques have enabled molecular definition of cortical laminae in the human brain beyond classic histological definitions [4,5], topographically-organized regions of the noradrenergic locus coeruleus [6], and gestational stages of the human fetal brain [7]. Findings from these studies highlight the importance of retaining spatial context in molecular profiling for biological insight.

The structure and function of the hippocampal formation (HPC), which is highly conserved across species, supports many fundamental processes, including learning, memory and stress response [8,9]. The HPC is composed of the dentate gyrus (DG), notable for its dense granule cell layer (GCL), and the cornu ammonis (CA), subdivided into CA1-4 regions, each displaying a distinct laminar organization and specialized cell types. Axons and dendrites from HPC neurons are organized in neuropil-enriched layers with sparse somata, such as DG molecular layer (ML), stratum lucidum (SL), and stratum radiatum (SR). Key to HPC circuit architecture is the trisynaptic loop, wherein entorhinal cortex (EC) inputs traverse from DG to CA3 to CA1. The circuit culminates with a relay to the subiculum (SUB), which sends projections to diverse regions as a major HPC output nucleus [10,11]. HPC cell types [12,13] have been well studied at the molecular and circuit level in rodent models, but the extent of conservation in human HPC is not well understood. While single-nucleus RNA-sequencing (snRNA-seq) has provided some insight [14,15], these studies lack spatial resolution, and do not retain cytosolic or synaptic transcripts [16]. Given its intricate spatial topography, and substantial local translation in HPC cell types, these constraints are critical.

To address these limitations, we generated paired SRT and snRNA-seq data to investigate spatial gene expression at cellular resolution in the adult human HPC. We integrated these datasets using spot-level deconvolution and non-negative matrix factorization (NMF) to illustrate complex molecular organization of the HPC across cell types, cell states, and spatial domains. We also identify cell populations and gene expression patterns in human HPC that correspond to mechanistically relevant patterns of neural activity and circuit connectivity in mouse HPC. Our findings underscore the value of SRT for revealing structure-function relationships in the brain across species, and provide novel biological insights about how molecular signaling and genetic risk in the human HPC is implicated in complex brain disorders.

2 | Results

2.1 | Experimental design and overview of profiling human hippocampus

We obtained tissue blocks of postmortem human brain containing anterior HPC from 10 neurotypical adult brain donors (**Fig. 1A, Supplementary Table 1**), and used histological staining to determine neuroanatomical orientation and inclusion of major subfields. Then, we used the 10x Genomics Visium Spatial Gene Expression (Visium-H&E) and 3' Single Cell Gene Expression platforms to generate paired SRT maps and snRNA-seq data of the human HPC. For each donor, we first collected 1-2 100µm cryosections for snRNA-seq, and then mounted 10µm cryosections across multiple Visium capture areas to profile all major subfields (CA1-4, DG, SUB)**.** We performed Visium-H&E for *n*=36 capture areas (2-5 per donor) (**Fig. 1B, Extended Data Fig. 1A**). We visualized cytoarchitecture from the H&E images and confirmed expected gene expression patterns of *SNAP25* and *MBP* across HPC structure (**Extended Data Fig. 1A**).

We performed a second SRT experiment using the Visium Spatial Proteogenomics (Visium-SPG) platform on two donors (*n*=8 total capture areas). Visium-SPG replaces H&E histology with immunofluorescence (IF) staining, enabling labeling of 4 broad cell types, including NeuN (neurons), OLIG2 (oligodendrocytes), GFAP (astrocytes), and TMEM119 (microglia) with transcriptome-wide gene expression profiling (**Fig. 1C**, **Extended Data Fig. 1B**). We applied standard preprocessing and quality control workflows to all *n*=44 capture areas (Visium-H&E and Visium-SPG) to remove low-quality spots (**Extended Data Fig. 2**, **Extended Data Fig. 3**, **Extended Data Fig. 4**). We retained 188,762 spots – 150,917 from *n*=36 Visium-H&E capture areas and 37,945 from *n*=8 Visium-SPG capture areas. Due to differences between experimental protocols for Visium-H&E and Visium-SPG, we observed substantial differences in gene expression between Visium-H&E and Visium-SPG samples (**Extended Data Fig. 5**). Hence, to limit the impact of technical differences on domain identity, we used spatial domain detection outputs for the H&E samples from all donors for downstream differential expression and other analyses.

Following tissue collection for Visium experiments, we collected a second set of 1-2 100µm cryosections from each donor, which were pooled with the cryosections that were collected earlier for snRNA-seq (**Fig. 1D**). Nuclei were incubated with propidium iodide (PI) and anti-NeuN antibody to label intact nuclei and neuronal nuclei, respectively. We sorted 9,000 nuclei from each donor on PI+ fluorescence, followed by an additional 9,000 nuclei on both PI+ and NeuN+ fluorescence to facilitate enrichment of neuronal nuclei (**Methods**). Following sequencing, we applied standard processing and quality control (QC) workflows to remove empty droplets, doublets, and poor-quality nuclei over multiple rounds of QC (**Extended Data Fig. 6**, **Extended Data Fig. 7**, **Extended Data Fig. 8**, **Extended Data Fig. 9**, **Extended Data Fig. 10**, **Methods**). Due to poor nuclei yields following QC (**Extended Data Fig. 7**), we repeated data collection from three donors to ensure representation of all donors in the data. After processing, we retained 75,411 high-quality nuclei across all donors.

2.2 | Molecular profiles for spatial domains in the human hippocampus using spatially-resolved transcriptomics

To define spatial domains in Visium-H&E data, we used a data-driven, multi-sample workflow consisting of unsupervised spatial domain detection with `PRECAST` [17] using the top 2000 spatially variable genes (SVGs) identified by `nnSVG` [18] (**Methods**, **Extended Data Fig. 11**, **Supplementary Table 3**). We considered a range of spatial domain resolutions (*k)* and selected *k*=18 based on Akaike information criterion (AIC) (**Extended Data Fig. 12A**).

Data-driven spatial domains were annotated based on inclusion of known anatomical subregions and expression of canonical marker genes (**Fig. 2A**, **Extended Data Fig. 13**). Due to similar gene expression, domains corresponding to CA2-4 and CA1 were collapsed into two domains, respectively (**Extended Data Fig. 12B-D**). Clustering identified domains mapping to DG layers, including the granule cell layer (GCL) and molecular layer (ML). We also annotated domains corresponding to pyramidal cell regions in HPC (CA2-4 and CA1), subiculum (SUB), subiculum/retrohippocampal area (SUB.RHP) and retrohippocampal area (RHP). We identified a sparsely distributed domain enriched for GABAergic neuron markers. Other domains mapped to HPC neuropil-enriched subregions, including stratum lucidum (SL) and stratum radiatum (SR). We identified a spatial domain mapping to neuropil-rich regions including both stratum lacunosum-moleculare (SLM) and DG subgranular zone (SGZ) (**Extended Data Fig. 12C**,**E, Supplementary Table 2**). Remaining domains mapped

to WM-rich regions, choroid plexus (CP), or vasculature (**Fig. 2A**, **Extended Data Fig. 13**). Based on anatomy and gene expression, we identified spatially-restricted spots corresponding with thalamus [19,20] and amygdala [21–23] (**Methods**, **Extended Data Fig. 14**), which were excluded from differential expression (DE) analysis. Neuronal and CP domains had higher library sizes, numbers of detected genes, and cellular density compared with other clusters (**Extended Data Fig. 15A-B**, **Extended Data Fig. 16**). Conversely, the percent of mitochondrial expression was higher and nuclei density lower in neuropil-enriched domains (**Extended Data Fig. 15C**). Separately, we also used histological staining and expression of known marker genes to manually label anatomically-defined HPC subfields [10,14,15,24] (**Extended Data Fig. 17**). Manual annotations and `PRECAST` domains were strongly correlated in some regions (e.g., GCL), but lower for others (e.g., SR.SLM) (**Extended Data Fig. 18**). `PRECAST` results were consistent with those generated from a second domain detection algorithm using graph-based autoencoders, `GraphST` [25], demonstrating robustness of spatial domain predictions across computational algorithms (**Extended Data Fig. 19**).

Next, we asked if we could discern transcriptomic profiles for HPC spatial domains using DE analysis. We first pseudobulked data by capture area, aggregating all UMIs within all spots for each domain (**Fig. 2B**). Principal components analysis (PCA) applied to pseudobulked data revealed that top components of variation corresponded with biologically meaningful differences between spatial domains (**Extended Data Fig. 20, Extended Data Fig. 21, Extended Data Fig. 22**, **Extended Data Fig. 23**). More of the variance was explained by spatial domain (**Extended Data Fig. 24**) compared with other covariates such as capture area, donor, Visium slide, sex, or age. To capture transcriptional differences between broader-level tissue structures, we also grouped domains into four broad domains based on data-driven gene expression similarities and known HPC anatomy: neuron-enriched domains, neuropil-enriched domains, WM, and vascular/cerebrospinal fluid (CSF)-related domains (**Fig. 2C**).

To identify differentially expressed genes (DEGs) across the fine-grained spatial domains, we employed a 'layer-enriched' linear mixed-effects modeling strategy to test for differences between one domain versus all others (adjusted for sex, age, and sequencing batch) as previously described [4,5] (**Extended Data Fig. 25**, **Extended Data Fig. 26**, **Supplementary Table 4**). We used the results from this enrichment model to identify marker genes for spatial domains (**Fig. 2D**). We confirmed canonical marker genes for spatial domains, such as *PPFIA2* for GCL (**Fig. 2E**), *PRKCG* for pyramidal neuron layers CA1-4 (**Fig. 2F**), and *MOBP* for WM. We also identified several novel marker genes, particularly with respect to HPC neuropil-enriched domains. These included *APOC1*, enriched in ML, SR, and SL (**Fig. 2G**), and *SFRP2*, which marks SLM/SGZ (**Fig. 2H**).

Using the same modeling strategy, we also asked if we could identify genes specific to broad domains (**Supplementary Table 5**). This analysis identified genes enriched in each of the 4 broad domains (**Fig. 2I**), and revealed marker genes that were enriched across all spatial domains within each broad domain, including *CLSTN3* for neuron-rich areas, *SLC1A3* for neuropil-enriched domains, *SHTN1* for WM domains, and *TPM2* for vascular/CSF domains (**Fig. 2J**). Altogether, these findings reveal widespread differences in spatial gene expression across the HPC corresponding with both discrete subregions and broad domains.

2.3 | Molecular profiles for cell types in the human hippocampus using paired single-nucleus RNA-sequencing

We used snRNA-seq data from all ten donors to define human HPC cell types (**Fig. 1D**). After applying QC and preprocessing steps (**Extended Data Fig. 6**, **Extended Data Fig. 7**, **Extended Data Fig. 8**, **Extended Data Fig. 9**, **Extended Data Fig. 10**, **Extended Data Fig. 27**), we applied graph-based clustering and identified

k=60 fine-grained clusters (annotated as “superfine” cell classes) from 75,411 nuclei. Next, we grouped the superfine cell classes into 24 “fine” cell classes based on canonical marker genes [10,14,15,24] (**Fig. 3A-B, Supplementary Table 2**). These classes included neuronal cell types (e.g. granule cells, or GCs), glia cell types (e.g. oligodendrocytes), and vascular or CSF-related cell types (**Fig. 3A**). Similar to SRT data, we found a small number of neurons from adjacent thalamus [20] and amygdala (**Methods**, **Extended Data Fig. 28**). Thalamic and amygdalar neurons were excluded from differential expression analyses to avoid obscuring within-HPC marker genes, but were included in the UMAP embedding and public data release. Finally, we aggregated (**Methods, Extended Data Fig. 29**) the “fine” cell class labels (*k*=24, **Fig. 3B, Extended Data Fig. 30A**) to lower resolution, including “mid” cell class labels (*k*=8, **Extended Data Fig. 31B**) and “broad” cell class labels (*k*=5, **Extended Data Fig. 31C**). Proportions of nuclei varied across donors with respect to all classification levels. As expected, neurons were enriched in PI+NeuN+ (85% of all nuclei) compared to PI+ (47% of all nuclei) sorted samples (**Extended Data Fig. 31**).

Using pairwise *t*-tests by gene, we found marker genes distinguishing “mid” cell class labels (**Supplementary Table 9**). These included genes highly expressed across all fine cell classes included within each mid cell class label (**Fig. 3C**), such as *GAD2* for all GABAergic neurons and *MOBP* for oligodendrocytes. Other marker genes reflected specific cellular functions. For instance, both ependymal cells and choroid plexus cells showed enrichment of genes related to motile cilia, such as *DNAH11*, which circulate CSF through ventricles. We asked if we could apply the same approach at the fine (**Supplementary Table 7**) and superfine cell class levels (**Supplementary Table 8**) to identify more fine-grained marker genes. These analyses revealed distinct molecular profiles for superfine and fine cell classes (**Fig. 3D**). For instance, we annotated different GABAergic neuron subtypes based on expression of classical markers, such as *SST*, *LAMP5*, *HTR3A*, and *VIP*. Pyramidal RHP neurons could be identified based on expression of canonical marker *SATB2* and classified into putative layers based on markers such as *TLE4* (layers 6/6b) and *CUX2* (layers 2/3). We also identified novel markers, like *ESR1* for hippocampus-amygdala transition area (HATA) neurons.

We used multiple strategies to validate both our annotations and their marker genes to increase confidence in these results. We verified marker genes identified with pairwise *t*-tests using the same pseudobulk linear modeling methods used for DE analysis of the SRT data, once again aggregating by sample and including covariates for sex, age, and sequencing batch (**Supplementary Table 9**, **Supplementary Table 10**, **Supplementary Table 11**). We also compared our annotations against a recent human HPC snRNA-seq dataset [15] by using gene-level *t*-statistics for each superfine cell class label from the pseudobulk enrichment model. We found high correlation where expected – for example between the same CA pyramidal cell types – along with low levels of correlation for some types unique to this dataset, such as choroid plexus cells and HATA neurons (**Extended Data Fig. 31**). Taken together, these results highlight the complex cellular architecture of the human HPC.

2.4 | Identifying cell type composition differences across human hippocampus using spot-level deconvolution

To spatially register cell classes to spatial domains, we correlated gene-level enrichment model *t*-statistics for each superfine cell class label and each fine-grained spatial domain. Gene expression profiles for some cell classes and spatial domains overlapped strongly (e.g. GCs with GCL; choroid plexus cells with choroid plexus domain), while other domains contained multiple cell classes (e.g., GABAergic cell types correlated strongly with the GABA domain, **Extended Data Fig. 32**). Unsurprisingly, we observed variable nuclei density across the data-driven spatial domains. For example, spots contained an average of 5 cells, ranging from 2.7 in the cell-sparse ML to 9.7 in the cell-dense GCL (**Extended Data Fig. 16**). Given this variation and the diversity of

cellular populations in the snRNA-seq data (**Fig. 3**), we explored cell type composition across spatial domains using three spot-level deconvolution algorithms [26]: `cell2location` [27], `RCTD` [28], and `Tangram` [29]. First, we used the Visium-SPG data where IF staining tagged four broad cell types: neurons (NeuN), oligodendrocytes (OLIG2), astrocytes (GFAP), and microglia (TMEM119) (**Fig. 1C, Extended Data Fig. 1B**) to evaluate these spot-level deconvolution algorithms by comparing known cell composition to predicted cell composition within each spot [5].

We considered both the fine and mid cell classifications from the snRNA-seq data (**Fig. 3B-C**, **Extended Data Fig. 29**). We identified the top gene markers for each fine cell class using the Mean Ratio method [30] (**Extended Data Fig. 33**), and confirmed that detected marker genes were specific to each cell type by comparing expression of the top genes for each cell class to canonical gene markers in the expected spatial domains (**Fig. 4A**, **Extended Data Fig. 34**, **Extended Data Fig. 35**, **Supplementary Table 12**). `RCTD` dropped a high number of spots using the Mean Ratio marker genes compared to `cell2location` and `Tangram`, thus, the default marker gene detection strategy was employed (**Methods**, **Extended Data Fig. 36**).

We then benchmarked `cell2location`, `RCTD`, and `Tangram` by calculating the predicted cell type composition within each spot in the Visium-SPG data (**Fig. 4A, Fig. 4B(i)**). We collapsed fine and mid cell classes into the broad class (**Extended Data Fig. 29**), which represented the IF staining: neurons, astrocytes, microglia, oligodendrocytes, and other (which was comprised of OPCs, vasculature, and CSF-related cells). To classify each nucleus in the Visium-SPG images as one of the broad cell types, we employed a classification and regression tree (CART) model to predict cell types based on the IF (**Fig. 4Bii**, **Extended Data Fig. 37**).

We then quantified algorithm performance using 2 complementary approaches: 1) comparing predicted cell type counts to those calculated from IF images (**Fig. 4B**, **Extended Data Fig. 38**) and 2) evaluating the localization of cell types to their expected spatial domain in the Visium-SPG and Visium-H&E data (**Fig. 4D**). Using the first approach, predicted counts from `RCTD` and `cell2location` had the highest correlation to CART-calculated counts (**Extended Data Fig. 38**, **Extended Data Fig. 39**, **Supplementary Table 13**). `RCTD` showed the most accurate and consistent performance at both mid and fine resolutions across all cell types and samples (**Fig. 4B(iii),C**; **Extended Data Fig. 39**). For the second approach, we first used a transfer learning approach to predict fine-grained spatial domain membership for all Visium-SPG samples (**Methods**, **Extended Data Fig. 40**). As expected, all 3 tools (`cell2location, RCTD` and `Tangram`) showed the highest predicted cell type proportion for GC excitatory neurons in spots clustered as GCL and ML (**Fig. 4D(i)**) compared to other domains. `RCTD` and `cell2location` performed well across all cell types, but `Tangram` failed to accurately map most cell types to the correct spatial domain (**Extended Data Fig. 41A**). After applying all three tools on the Visium-H&E data,(**Fig. 4D(ii)**), we noted that `RCTD` exhibited consistent predictions at both mid and fine levels, closely aligning with biological expectations and revealing spatial variations in cell types (**Extended Data Fig. 41B**), while `Tangram` and, to a lesser degree, `cell2location`, predicted similar cellular composition in each domain. In summary, these analyses confirmed `RCTD` is most suitable for spot deconvolution of Visium HPC data because it accounts for the cellular heterogeneity seen across the HPC.

2.5 | Non-negative matrix factorization (NMF) uncovers complex gene expression patterns shared between snRNA-seq and SRT data

Spot-level deconvolution methods are valuable, but most methods rely on *a priori* classification of data into discrete clusters, thereby overlooking transcriptional heterogeneity within cell types. Coupling non-negative matrix factorization (NMF) with transfer learning provides an intuitive approach to identify continuous gene

expression patterns that may be more functionally relevant [31]. NMF decomposes high-dimensional data into *k* lower-dimensional latent spaces [32] **(Methods**). When applied to snRNA-seq data, latent factors represent distinct gene expression patterns that can define biological or technical variation. Patterns may be specific to cell types, cell states, or other biological processes [33,34]. Transfer learning can map identified patterns onto independent datasets [34], allowing for discernment of patterns shared across species and data modalities.

Here, we used NMF to define gene expression patterns within the snRNA-seq data and transfer learning to probe expression of these patterns within the SRT data (**Fig. 5A**). Using the `RcppML` R package [32], we performed NMF on the normalized snRNA-seq counts matrix at rank *k*=100 to define 100 NMF patterns (**Extended Data Fig. 42**, **Supplementary Table 14**). To enable comparisons between the contributions of specific genes and nuclei to each pattern, pattern weights are normalized such that both nuclei-level and gene-level NMF weights are interpreted as proportions of a given pattern. We used transfer learning to predict spot-level weights for the 100 patterns in the SRT data to allow for spatial visualization and analysis of patterns (**Methods, Fig. 5A**).

To validate our approach, we asked if we could find NMF patterns specific to superfine cell classes (**Fig. 3A**). We identified patterns specific to 55 of 60 superfine cell classes (**Extended Data Fig. 43**). For instance, one pattern (nmf77) was strongly enriched in oligodendrocytes (**Fig. 5B-C**) and localized in oligodendrocyte-rich domains of SRT data, including the SLM and WM-enriched domains (**Fig. 5D**). We also identified patterns mapping to astrocytes (**Extended Data Fig. 44**), such as nmf79 (**Fig. 5E-F**), which is primarily localized to neuropil-rich domains in the SRT data (**Fig. 5G**). Some patterns mapped to more than one cell type, likely reflecting complex biological processes not specific to a single cell type. These included nmf13, primarily found in neuron cell classes (**Fig. 5H-I**) and neuron-enriched domains (**Fig. 5J**). We identified several patterns that represented technical variables and donor-specific artifacts (**Methods, Extended Data Fig. 45**, **Extended Data Fig. 46**) [35], which were excluded from marker gene detection and enrichment analyses.

We identified marker genes for all patterns using `patternMarkers` function from `CoGAPS` [36] (**Fig. 5J; Supplementary Table 15**). Marker genes for nmf77, such as *PLP1* and *MAG*, were consistent with oligodendrocyte identity. Marker genes for nmf79 included solute transporters, such as *SLC25A18* and glutamate transporter genes *SLC1A2* and *SLC1A3*, typically expressed in astrocytes. Marker genes for nmf13 included genes important in post-synaptic signaling, including synapse-associated genes (e.g. *PDZD2* and *DLGAP3*), potassium channels (e.g. *KCNQ2*), and signaling molecules (e.g. *RASAL1* and *PDE2A*). We performed over-representation analysis to identify Gene Ontology (GO) [37] terms enriched in marker genes for each pattern from all three ontologies. We identified at least one enriched GO term for 83 of 100 patterns (**Fig. 5K, Extended Data Fig. 47; Supplementary Table 16**). Unsurprisingly, top GO terms for oligodendrocyte-associated pattern nmf77 included terms related to oligodendrocyte function, confirming the efficacy of this strategy (**Fig. 5L**). Top terms for astrocyte-associated pattern nmf79 included terms related to metabolism and transmembrane transport (**Fig. 5L**). Top terms for other astrocyte-associated patterns highlighted other distinct processes important to astrocyte function, such as extracellular matrix production (**Extended Data Fig. 44**), suggesting that NMF patterns enable identification of multiple astrocyte phenotypes. Finally, enrichment for nmf13 included terms associated with synapse structure and function, suggesting that this pattern might reflect molecular pathways related to synaptic structure or plasticity (**Fig. 5L**). Overall, these results validate the ability of NMF and transfer learning to capture continuous gene expression patterns associated with cell class, biological processes, and technical variation, which can be transferred and examined across datasets.

2.6 | NMF and transfer learning enable spatial mapping and inference of circuit connectivity of pyramidal neurons in the hippocampal formation

Both HPC and RHP contain multiple different classes of pyramidal neurons with distinct molecular profiles [12,14,15], physiological properties [38], axonal projection targets [13], and spatial localization patterns [39–41]. Recent studies indicate a diversity of excitatory neuron types within rodent subiculum and RHP, each with different circuit connectivity patterns [10,11]. `PRECAST` identified canonical spatial domains harboring specialized neuron types, including CA1 and SUB. However, spatial domains may contain multiple spatially-organized excitatory neuron types. We asked if we could elucidate fine-grained spatial arrangement of these cells using NMF and transfer learning.

We identified 20 NMF patterns mapping to HPC and RHP pyramidal neurons in the snRNA-seq data (**Extended Data Fig. 43**). We also identified 3 patterns mapping to amygdala pyramidal neurons, which were excluded from this analysis (**Methods**). We compared distribution of these patterns within *a priori* superfine cell classes of pyramidal neurons defined by graph-based clustering (**Fig. 6A**). Each pattern showed enrichment in specific cell classes, with most mapping to 4 or fewer superfine cell classes. Following transfer of NMF patterns to SRT data, we also found enrichment of specific patterns in neuron cell body-rich spatial domains (**Fig. 6B**). Spatial domains also showed different levels of cell class diversity. For example, patterns mapping to the CA1 domain were mostly nmf15, specific to CA1 neurons, and nmf32, specific to CA1 and prosubiculum (ProS) neurons. By contrast, SUB.RHP and RHP domains harbored patterns associated with multiple cell classes. Pyramidal neuron patterns also had distinct marker genes (**Fig. 6C**). In the SRT data, patterns associated with CA1-4 pyramidal neurons spatially localized to CA1-4 subregions as expected (**Fig. 6D**). Patterns associated with prosubiculum, subiculum, and RHP pyramidal neurons showed a distinct laminar and compartmentalized pattern within SUB, SUB.RHP, and RHP spatial domains (**Fig. 6E**). In summary, NMF analyses revealed molecular signatures and nuanced spatial patterns of pyramidal neurons at better resolution than other methods.

HPC pyramidal neurons can also be classified by the brain region that their axons target. Neurons projecting to different target regions have unique functions and physiological characteristics [42]. A recent study in mice coupled retrograde viral circuit tracing with single-cell methylation sequencing (snmC-seq) to investigate whether gene-regulatory elements can define neuronal cell types based on their axonal projection targets [43]. The snmC-seq data was used to predict transcriptional profiles and classify the labeled neuronal projectors. To add putative circuit information to human data, where connectivity resolution is lacking, we capitalized on the snmC-seq mouse data to ask whether we could use NMF patterns to infer the projection targets of human HPC pyramidal neurons. We mapped NMF patterns from the human snRNA-seq data onto mouse snmC-seq data using the transfer learning approach described above (**Fig. 5A, Methods**). NMF patterns corresponding to pyramidal neurons mapped to mouse neurons with similar cell type labels (**Fig. 6F, Supplementary Table 2**). Additionally, NMF pyramidal neuron pattern weights from the human data showed a gradient across axonal projection targets in the mouse data (**Fig. 6G, Supplementary Table 2**). Patterns specific to CA1-4 and subiculum neurons were weighted higher in mouse neurons projecting to olfactory bulb, hypothalamus, entorhinal cortex, retrosplenial cortex, and thalamus. Patterns specific to RHP neurons were weighted higher in mouse neurons projecting to hippocampus, primary motor cortex, anterior cingulate cortex, piriform cortex, and prefrontal cortex. These signatures could be transferred across species and data modalities to reveal spatial arrangement of neuron classes and infer neuron projection targets in the human brain. Altogether, in addition to resolving molecular signatures of human pyramidal neuron populations, NMF and transfer learning enabled prediction of circuit connectivity in human HPC by leveraging mouse retrograde circuit tracing data.

2.7 | NMF identifies cell populations primed for activity-induced recruitment in the human hippocampus

The inability to test causality of gene expression associations in postmortem human tissue studies is a major limitation. However, integrating data from model systems that incorporate experimental manipulations can aid functional interpretation. To demonstrate this approach, we leveraged snRNA-seq data generated in a mouse model of electroconvulsive seizures (ECS) to identify cell type-specific activity-regulated gene expression patterns in the HPC [44]. We asked if we could identify cells more likely to be activated by stimulation and find gene expression patterns predicting induction of neural activity. Using transfer learning, we mapped NMF patterns derived from the human HPC snRNA-seq data to mouse snRNA-seq data of HPC neurons at baseline and neurons activated by induced ECS. We confirmed that patterns specific to neuron classes in the human HPC snRNA-seq data were enriched in similar groups of neurons in the mouse HPC data (**Extended Data Fig. 48**). We then asked if we could identify human NMF patterns specific to control or ECS-activated states in mouse HPC neurons (**Extended Data Fig. 49**). Some patterns correlated more strongly with unactivated mouse control neurons. These included nmf10, specifically enriched in control mouse GCs (**Fig. 7A**). In contrast, nmf14 correlated with ECS-activated mouse neurons and was enriched in activated mouse GCs (**Fig. 7B**). Other patterns more strongly associated with ECS activation included nmf91, nmf20, and nmf13, which were distributed across cell types (**Fig. 7C-E**). Surprisingly, nmf10 and nmf14 were specifically enriched in different GC superfine cell classes in the human snRNA-seq data. While nmf10 had higher weights in GC.1 and GC.2 cell classes (**Fig. 7F**), nmf14 was mostly enriched in GC.4 (**Fig. 7G)**.

Because nmf91, nmf20, and nmf13 did not map to a specific superfine cell class (**Extended Data Fig. 43**), we asked if these might represent different biological processes. Top GO terms for nmf91 were more related to response to stimuli (e.g. peptide hormone), sterol metabolism, and transcription factor activity. (**Supplementary Table 16**). However, enriched GO terms for nmf20 were associated with synaptic plasticity and MAPK signaling, while terms for nmf13 were related to synaptic structure and potassium channels (**Extended Data Fig. 47**). To further functional interpretation, we compared transcriptional signatures of each of these 5 NMF patterns (**Fig. 7H**). Marker genes for nmf91 included immediate early gene transcription factors such as *FOS, JUN,* and *NR4A3*. In contrast, nmf20 marker genes included synaptic plasticity-associated genes such as *SORCS3* [45] and *BDNF*. Finally, we asked whether these patterns showed spatial topography in the human HPC. In the human HPC SRT data, GC-associated patterns nmf10 and nmf14 primarily localized to GCL and were intermingled in mostly non-overlapping spots throughout GCL and ML (**Fig. 7I**). In contrast, nmf13 was distributed across neuron cell body-rich spatial domains. While nmf20 was mostly enriched in neuron cell body-rich domains, it was more sparse than nmf13. Very few spots expressed nmf91, which was sparsely scattered across all domains (**Fig. 7J**). Together, these results illustrate how NMF and transfer learning can be used for cross-species integration of mouse transcriptomics data that incorporates experimental manipulations with human SRT and snRNA-seq data to provide functional biological insight.

2.8 | Localization of genetic risk for brain disorders across the spatio-molecular topography of the human HPC

The HPC is implicated in many pathological conditions, although specific cell types and spatial domains driving these associations remain unclear [46–49]. To identify cell types, spatial domains, and gene expression patterns associated with genetic risk for diseases and disorders, we used stratified linkage disequilibrium score regression (S-LDSC) [50–52] to calculate enrichment of heritability of multiple polygenic traits, including

neuropsychiatric and non-psychiatric conditions [53–68]. S-LDSC regression coefficients represent contribution of a given annotation (e.g. spatial domain, cell class, or NMF pattern) to heritability of a given trait (**Methods**).

We first performed S-LDSC regressions for each spatial domain. We observed enrichment of genetic risk for schizophrenia (SCZ) in the CA1 domain, and multiple disorders, including SCZ, in the RHP (**Extended Data Fig. 50A**). Since spatial domains contain heterogeneous cell populations, we performed the same analysis for each superfine cell class. We observed many expected associations, including enrichment of Alzheimer's disease (AD) risk in microglia [54] (**Extended Data Fig. 50B**). To leverage the more fine-grained resolution afforded by the NMF, we also performed this analysis across NMF patterns (**Extended Data Fig. 50C**). We identified multiple patterns unique to various cell types and spatial domains enriched for neuropsychiatric disorder risk, including a pattern mapping to SST/CORT neurons (nmf74) enriched for bipolar disorder (BPD), a LAMP5+ neuron pattern (nmf35) enriched for BPD, major depressive disorder (MDD), and SCZ, and a VIP+ neuron pattern (nmf50) enriched for smoking initiation (**Fig. 8A**). Inhibitory synaptic signaling has been implicated in various psychiatric disorders, and subpopulations of GABAergic neurons are organized in specific subfields of the hippocampus [69–71]. We confirmed the specificity of these NMF patterns to cell types by calculating the distribution of pattern weights within subpopulations of GABAergic neurons (**Fig. 8B**). We then illustrate their spatial organization (**Fig. 8C**), specifically noting that VIP and LAMP5 interneurons are localized in neuropil-rich regions, while SST neurons are located in neuron-rich subfields, similar to expectations from mouse literature [70,71]. Our results illustrate that, although the HPC is implicated in various neuropsychiatric disorders, genetic risk for these disorders reside in different cell types and spatial domains.

2.9 | Data Access and Visualization

This atlas of integrated single cell and spatial transcriptomic data is the most comprehensive map of the human HPC to date. To enable exploration of these rich data resources, we created multiple interactive data portals, available at research.libd.org/spatial_hpc/. The pseudobulked snRNA-seq and SRT data are freely available through *iSEE* apps [72] to allow users to visualize expression for genes of interest across cell types and spatial domains. To facilitate exploration of SRT data in anatomical context, we merged all Visium capture areas from each donor for visualization of high resolution images and corresponding gene expression data, NMF patterns, and spot-level deconvolution results at the donor level using a web-based interactive tool, Samui Browser [73] (10 donors, *n*=36 Visium-H&E capture areas; 2 donors, *n*=8 Visium-SPG capture areas. Finally, raw and processed data are available through Gene Expression Omnibus (GEO) [74] under accession GSE264624 to facilitate access for methods development and in depth analysis.

3 | Discussion

By integrating Visium-H&E, Visium-SPG, and snRNA-seq data, we generated a comprehensive transcriptomic atlas of the adult human hippocampus (HPC). We characterized the molecular organization of the human HPC with both cellular and spatial resolution, identifying discrete spatial domains and a rich repertoire of HPC cell types. Non-negative matrix factorization (NMF) uncovered gene expression patterns representing both cell type-specific molecular signatures and shared biological processes, such as synaptic signaling (**Fig. 5**). By integrating these patterns with SRT data, cross-species functional genomics data (**Fig. 6**, **Fig. 7**), and GWAS summary statistics (**Fig. 8**), we provide spatial context to disease-relevant cell types, cell states, neural circuits, and molecular pathways.

Viral circuit labeling tools have revolutionized the ability to profile neuron populations in mouse models based on their innervation patterns. Using these tools, projection targets of mouse ventral HPC neurons have been topographically [75] and transcriptionally profiled [76]. Coupling circuit labeling with optogenetic or chemogenetic approaches enables selective manipulation to better understand circuit-specific functions. For example, molecularly distinct projections from ventral HPC to the medial prefrontal cortex mediate context-dependent fear memory retrieval [77], while inputs to the lateral hypothalamic area regulate avoidance behavior [78]. Understanding functional dynamics in defined hippocampal circuitry is important because dysregulation of activity in these circuits is implicated in various neuropsychiatric disorders [79,80]. Identifying unique molecular identities for topographically-organized HPC cell types based on their innervation targets in the human brain is critical for the goal of developing circuit-specific therapeutics. However, circuit tracing and manipulation are not possible in postmortem human brain tissue. To circumvent this issue, we capitalized on the power of molecular genetic tools to to extrapolate projection targets of molecularly-defined human HPC neuron types by leveraging mouse epigenomic and retrograde viral circuit tracing data [43] (**Fig. 6E-F**). Correspondence between specific molecular profiles and inferred target regions was consistent with known biology, highlighting the validity of our approach. For example, we identified a strong association between a RELN+ L2/3 human neuron signature (nmf84) and mouse L2/3 entorhinal cortex (EC) neurons projecting to HPC, likely representing EC L2 stellate cells [81]. We also identified specific cell types with putative connections to limbic regions including hypothalamus, prefrontal cortex, and anterior cingulate cortex, which are implicated in disorders such as schizophrenia [82] and depression [83,84]. Indeed, some therapeutic approaches, such as subcallosal cingulate deep brain stimulation, were developed with the goal of normalizing circuit-level dysfunction [85–87]. However, molecularly informed circuit definition in the human HPC has the potential to facilitate more precise manipulation of cells and circuits, thereby improving clinical outcomes.

Functional studies in rodents that can test causality of cellular and molecular associations are powerful, but they lack direct relevance to human brain function and behavior. Conversely, measuring transcriptomic changes in response to experimental manipulations is not possible in human postmortem brain tissue. However, integrating human transcriptomic data with functional genomics data in animal models has the potential to facilitate interpretation of cellular responses to various stimuli. To illustrate the feasibility of this approach, we mapped NMF patterns from the human HPC onto snRNA-seq data from a mouse model of induced electroconvulsive seizures (ECS) [44] to identify gene expression patterns that are putatively associated with neural activity in the human brain (**Fig. 7**). Identifying and localizing expression patterns associated with activity-regulated genes (ARG) in the human HPC is important because expression of these genes is critical for recruitment of HPC neuronal ensembles controlling learning, memory and other cognitive processes [88]. NMF patterns that reflected processes known to be regulated by activity, including immediate early gene expression and synaptic signaling, were enriched in activated mouse neurons (**Fig. 7A-E**). These patterns were distributed across multiple neuron types and spatial domains in human HPC, including patterns in human GCs that were differentially associated with either control or activated neurons (**Fig. 7F-G**). These patterns were sparsely distributed in non-overlapping GCL spots in SRT data (**Fig. 7J-K**). In the rodent, the majority of GCs are relatively silent under both baseline conditions [89] and during exploration of novel environments [90,91]. However, a small fraction of these GCs rapidly increase their firing activity during behavior, supporting a sparse coding scheme that facilitates precision in both pattern and novelty detection during learning [92]. NMF patterns that differentially map to control versus activated neurons may represent the population of GCs that are more quiescent versus those that are either actively firing or intrinsically primed for activity [93]. Indeed, active and quiescent mouse GCs have substantially different transcriptomic profiles [44], and differences in GC activity levels may reflect differences in their ability to be recruited into neuronal ensembles in response to behavioral experiences [94–99]. These cross-species analyses provide insight about the putative spatial and molecular organization of cellular activity states in the human HPC, which is critical for

attempting to design rational therapies for disorders involving dysregulated HPC neuron activity. For instance, HPC neuron hyperexcitability is a known component of Alzheimer's disease pathology [100–103]. Future studies can also leverage these human ARG patterns to design complementary studies in animal models to explore transcriptional signatures associated with learning, memory, and cognition.

In summary, we demonstrate that NMF is an effective approach to integrate transcriptomics data with other datasets, particularly across species and data modalities. Our NMF-based approach has some limitations. Namely, NMF can be sensitive to initialization values, patterns corresponding with noise may be incorrectly attributed to distinct biological processes, and NMF weights can be scaled differently across different factorizations. Despite these limitations, this approach can be effectively used to map other, external data to our atlas as illustrated. To facilitate wide access to our data, we provided the human HPC snRNA-seq and SRT atlas as a resource to the scientific community through multiple avenues, including interactive web applications for visualization and exploration.

4 | Methods

4.1 | Postmortem human tissue samples

Postmortem human brain tissue from neurotypical adult donors of European ancestry were obtained at the time of autopsy following informed consent from legal next-of-kin, through the Maryland Department of Health IRB protocol #12–24, and from the Western Michigan University Homer Stryker MD School of Medicine, Department of Pathology, the Department of Pathology, University of North Dakota School of Medicine and Health Sciences, and the County of Santa Clara Medical Examiner-Coroner Office in San Jose, CA, all under the WCG protocol #20111080. Using a standardized strategy, all donors underwent a comprehensive retrospective clinical diagnostic review to exclude for any lifetime history of psychiatric or substance use disorders. Macro- and microscopic neuropathological examinations were performed, and subjects with evidence of neuropathological abnormalities were excluded. Additional details regarding tissue acquisition, processing, dissection, clinical characterization, diagnose, neuropathological examination, RNA extraction and quality control (QC) measures have been previously published [104]. Demographic information for all donors in this study is listed in **Supplementary Table 1**.

4.2 | Tissue processing and quality control

All hippocampus dissections in this study were performed by the same neuroanatomist (co-author TMK). The hippocampus was dissected as uniformly as possible using anatomical landmarks to include the anterior portion of the hippocampus proper plus the subicular complex [49]. Frozen samples were mounted in OCT (TissueTek Sakura) and cryosectioned at -12°C (Leica CM3050s). To ensure high-quality samples, we measured RNA Integrity Number (RIN) in the sample blocks from a subset of donors and ensured that RIN in our dissected tissue samples was comparable to RIN calculated at time of brain donation (**Supplementary Table 1**). 10µm sections were mounted on standard microscopy slides for staining with hematoxylin and eosin (H&E) for orientation and quality control. Following QC, 2-4 100µm cryosections, totalling approximately 50mg of tissue, were transferred into a low-adhesion tube (Eppendorf) kept on dry ice and reserved from the anterior portion of the HPC for snRNA-seq. Then, two 10µm sections were mounted on standard microscopy slides and reserved. The tissue block was then scored using a razor blade, so that tissue sections approximately the size of a Visium capture area could be positioned and placed onto chilled Visium Spatial Gene Expression slides (part number 2000233, 10x Genomics). Following the successful completion of

Visium experiments, additional 100µm cryosections were collected for single nucleus RNAseq. In some cases, multiple Visium experiments were performed to ensure inclusion of relevant subfields. In these cases, distance along the anterior-posterior axis between experiments did not exceed 500µm. To assist in the reassembly of scored tissue sections, H&E staining was performed on one slide adjacent to the scored Visium sections. RNAscope (ACD) was performed on the other to guide histological annotation of canonical subfields, using previously defined markers of dentate gyrus (*PROX1*), CA3 (*NECAB1*), CA1 (*MPPED1*) and SUB (*SLC17A6*).

4.3 | Spatially resolved transcriptomics (SRT) data generation

Visium-H&E Data Generation

Visium Spatial Gene Expression slides were processed as previously described [4,5]. To ensure the optimal time duration of exposure to permeabilization enzyme, tissue optimization experiments were performed according to the manufacturer's protocols (protocol CG000160, revision B, 10x Genomics). Tissue sections were scored to include the dentate gyrus to facilitate orientation, and exposed to permeabilization enzyme for differing time durations. cDNA synthesis was performed using a fluorescently-labeled nucleotide (CG000238, revision D, 10x Genomics). The slide was then coverslipped and fluorescent images were acquired at 10x magnification with a TRITC filter (ex 550nm/em 600nm) on a Cytation C10 Confocal Imaging Reader (Agilent). Following this experiment, 18 minutes was selected as the optimal permeabilization time. For each Visium slide, H&E staining was performed (protocol CG000160, revision B, 10x Genomics), after which slides were coverslipped and high-resolution, brightfield images were acquired on a Leica CS2 slide scanner equipped with a 20x/0.75NA objective and a 2x doubler.

Visium-SPG Tissue Processing and Image acquisition

To enable labeling with four cell-type marker proteins (NeuN for neurons, TMEM119 for microglia, GFAP for astrocytes, and OLIG2 for oligodendrocytes), we performed combinatorial immunofluorescence staining paired with spatial gene expression using the Visium-SPG protocol (protocol CG000312, revision D, 10x Genomics) with some modifications. Briefly, gene expression slides were fixed in methanol, incubated in blocking buffer (3X SSC supplemented with 2% BSA, 0.1% Triton X100 and 2U/ul RNAse inhibitor (RiboLock, Thermofisher, Cat# EO0384)), then incubated with primary antibodies for 30 minutes at room temperature: mouse anti-NeuN antibody conjugated to Alexa 488 (Sigma Aldrich, Cat# MAB377X, 1:100), rabbit anti-TMEM119 antibody (Sigma Aldrich, Cat# HPA051870, 1:20), rat anti-GFAP antibody (Thermofisher, Cat# 13-0300, 1:100), and goat anti-OLIG2 antibody (R&D systems, Cat# AF2418, 1:20). Slides were washed with wash buffer (3X SSC with 2% BSA, 0.1% Triton X100, 2U/ul RNAse inhibitor and supplemented with ribonucleoside vanadyl complex (Concentration, NEB S1402S), then incubated for 30 minutes with secondary antibodies at RT: donkey anti-rabbit IgG conjugated to Alexa 555 (Thermofisher, Cat# A-31572, 1:300), donkey anti-rat IgG conjugated to Alexa 594 (Thermofisher, Cat# A-21209, 1:600), and donkey anti-goat IgG conjugated to Alexa 647 (Thermofisher, Cat# A-21447, 1:400). DAPI (Thermofisher, Cat# D1306, 1:3000, Final 1.67 µg/ml) was included as a nuclear counterstain. The slide was coverslipped with glycerol containing RNase Inhibitor (2 U/µl) and scanned on a Vectra Polaris slide scanner (Akoya Biosciences) at 20x magnification with the following exposure time per given channel: 2.70 msec for DAPI; 185 msec for Opal 520; 900 msec for Opal 570; 160 msec for Opal 620; 1300 msec for Opal 690; 100 msec for autofluorescence.

Visium Data Generation

Following removal of the coverslips, tissue was permeabilized, cDNA synthesis was performed, and sequencing libraries were generated for all Visium samples following the manufacturer's protocol (CG000239, revision C, 10x Genomics). Libraries were loaded at 300 pM and sequenced on a NovaSeq System (Illumina)

at the Johns Hopkins Single Cell Transcriptomics core according to manufacturer's instructions at a minimum depth of 60000 reads per spot.

4.4 | SRT data processing and analysis

Visium raw data processing

FASTQ and image data were pre-processed with the 10x `SpaceRanger` pipeline (version 1.3.1 for Visium-H&E and version 2.0.0 for Visium-SPG) [105]. Reads were aligned to reference genome GRCh38. For analysis steps, we used R (version 4.3.2, unless otherwise noted) and Bioconductor (version 3.17) for analysis of genomics data [106]. Outputs from the `SpaceRanger` pipeline were read into R and stored in a `SpatialExperiment` [107] (SPE) object using the `read10xVisiumWrapper()` from the `spatialLIBD` package [108]. For each donor, we first rearranged the order of capture areas. This was done such that overlapping or adjacent capture areas were adjacent to one another when plotted on a grid (2x2 grid for donors with 3 or 4 capture areas, 1x2 or 2x1 grid for donors with 2 capture areas, or 3x2 grid for donors with 5 capture areas). For donors with both Visium-H&E or Visium-SPG capture areas, capture areas were rearranged separately. After rearrangement, images from each capture area and spatial coordinates in the SPE were rotated to ensure that neighboring regions from each donor were spatially adjacent and correctly oriented to recapitulate the organization of the initial, unscored cryosection as well as possible. We performed these rotations with the `imgData()` function from the `SpatialExperiment` package. For a given H&E or fluorescent image, `imgData()` can transform the raster representation of the image data by some user-specified rotation values (in degrees). In our SPE object, we also independently rotated both the pixel coordinates for each spot (stored in the `spatialCoords` slot) and the capture area array coordinates for each spot (stored in `colData` slot as `array_row` and `array_col`). These rotations were performed to correspond with image rotations.

Visium quality control and count normalization

We filtered the data to remove all undetected genes and spots with zero counts. We used the `addPerCellQC()` function from the `scuttle` Bioconductor package to compute and store standard quality control metrics, such as mitochondrial expression rate, library size, and number of detected genes [109,110]. We applied a 3x median absolute deviation (MAD) threshold to discard spots with low library sizes and/or low numbers of detected genes (**Extended Data Fig. 2**; **Extended Data Fig. 3**). Mitochondrial expression rate was not used to define low-quality spots as this metric appeared to correlate with true biological variation (**Extended Data Fig. 4**). Specifically, hippocampal neuropil-enriched layers, such as dentate gyrus molecular layer, stratum radiatum, stratum lucidum, and stratum lacunosum-moleculare, showed enrichment of mitochondrial genes. These regions are characterized by large numbers of synapses and a paucity of neuronal cell bodies. In neurons, mitochondria are known to be enriched at both presynaptic and postsynaptic sites [111,112]. After applying these QC steps, the number of genes x the number of spots was 31,483 x 188,762. Counts were normalized by library size and $\log_2$-transformed using `computeSumFactors()` from `scran [109]` and `logNormCounts()` from `scuttle [110]`.

Feature selection, highly variable genes

For initial feature selection, we identified highly-variable genes (HVGs). Specifically, we first used `modelGeneVar()` from `scran` to model the mean-variance relationship of $\log_2$ normalized gene expression values for each gene [109]. We selected the top genes by biological component of variance (e.g. distance from the expected gene expression value based on the mean-variance relationship trend) as HVGs (1,985 genes), using the argument `prop=0.1` in the `getTopHVGs()` function from `scran`.

Dimensionality reduction

We used `runPCA()` from `scran` to perform principal component analysis (PCA) on the top HVGs. We retained the top 100 PCs. For visualization purposes, we used `runUMAP()` from `scran` to generate a low-dimensional uniform manifold approximation (UMAP) representation of the high-dimensional data [113]. We identified a batch effect between Visium-H&E and Visium-SPG samples (**Extended Data Fig. 5**). Because of this, we elected to analyze Visium-H&E samples separately to identify unsupervised clusters.

Feature selection, spatially variable genes

Highly variable genes do not necessarily capture spatial variance. Thus, for feature selection prior to unsupervised clustering of SRT data, we used the `nnSVG` Bioconductor package [18] to select the top 2,000 spatially variable genes (SVGs), or those genes with the most variance in expression across a tissue [114]. The `nnSVG` package uses a nearest-neighbor Gaussian process model. This approach offers linear scalability and efficient handling of large datasets. The model aims to identify genes with varying expression patterns across different regions within the tissue by estimating gene-specific spatial ranges and retaining detailed spatial correlation information.

Before SVG identification, we created a new SPE object without Visium-SPG samples. We filtered genes with fewer than 100 counts across all spots across all capture areas. For each capture area, we also filtered genes with less than 3 counts in at least 0.5% of spots in that capture area. We did not filter mitochondrial genes. After filtering, we ran `nnSVG` separately on each capture area using the default parameters. After running nnSVG for each sample, we ranked all genes within each capture area by spatial variance. We averaged gene ranks across all capture areas and ranked genes by mean rank. We retained the top 2,000 genes for clustering analysis (~10% of all genes retained after filtering steps), constraining to genes ranked in the top 1,000 genes in at least 2 different capture areas (**Extended Data Fig. 11**).

Unsupervised clustering of spatial transcriptomics data

For spatially-aware unsupervised clustering of SRT data, we used `PRECAST` v1.5 [17] for spatial transcriptomics data integration and clustering across multi-sample datasets. Briefly, PRECAST factorizes normalized gene expression matrices from each sample into a latent factor. The factorization approach aims to capture underlying patterns within each sample with a shared distribution across spatial clusters. Spatial dimension reduction, spatial clustering, and alignment of joint embeddings for biological effects are also performed across samples. An intrinsic conditional autoregressive (CAR) component is used to capture the spatial relationships between neighboring microenvironments or tissue regions. Additionally, a Potts model promotes spatial smoothness across clusters. `PRECAST` also performs batch correction internally. This is done through a projection approach, which is similar to deep learning methods. This approach accounts for both technical and biological differences, which allows for alignment of clusters across samples. We refer to `PRECAST` clusters as spatial domains throughout the text.

Prior to running `PRECAST`, we first split our `SpatialExperiment` object into a list of 10 objects, one for each donor. Spatial coordinates for each capture area within each donor were offset such that similar HPC subregions between capture areas were adjacent. The list of donor-specific `SpatialExperiment` objects was converted to a list of `Seurat` [115] objects, which `PRECAST` takes as input, using `Seurat` v4.3.0.

We ran `PRECAST` across all donors using *k*=5 through *k*=20. We used the top 2,000 SVGs calculated using `nnSVG` instead of the default highly variable genes (HVGs). Following `PRECAST`, we extracted BIC values from each `PRECASTObj`. We plotted BIC values for all *k* values. BIC decreased dramatically from *k*=5 to *k*=10, then decreased only slightly between each successive *k* value. We prioritized *k*=15 through *k*=20 as these had the lowest BIC values (**Extended Data Fig. 12A**). We visualized spatial domain membership for

k=15 through *k*=20 and found similar results. However, *k*=18 included a spatial domain that appeared to map to SLM and SGZ, which other clusterings lacked (**Extended Data Fig. 12B**). This domain also had high expression of astrocyte-specific markers (**Extended Data Fig. 12C**). Thus, for downstream analyses, we used spatial domains detected using *k*=18.

We also developed a multi-sample integration pipeline using `GraphST` [25] and `Harmony` [116] to obtain consistent clustering results while addressing batch effects. The first step is preprocessing the datasets to select the most variable genes (HVGs) that are common across all slides after QCing all spots. Then `GraphST` is employed to reconstruct the gene expression profiles for each individual slide based on the default tutorial. `Harmony` is applied to the concatenated gene expression reconstruction across samples to generate corrected top 30 Principal Components (PCs) embeddings, which take into account three specific covariates: the platform used for gene expression measurement (Visium-SPG or H&E), donors, and capture areas. Default parameters in `Harmony` are utilized. Finally, the corrected PCA components are used with `mclust`, a model-based clustering algorithm, to detect consistent spatial domains across the different slides.

Spatial domain annotation

We annotated spatial domains in a nested structure at two levels (domain and broad domain) based on anatomical location. Domains mapping to GCL, Cornu Ammonis pyramidal cell layers (CA2-4.1, CA2-4.2, CA1.1, CA1.2), subiculum (SUB), subiculum and retrohippocampal region (SUB.RHP), retrohippocampal region (RHP), and GABAergic neuron-rich spots (GABA) were grouped as neuron cell body-rich domains at the broad domain level. Domains mapping to dentate gyrus molecular layer (ML), stratum lucidum/stratum radiatum (SL/SR), stratum radiatum/stratum lacunosum-moleculare (SR/SLM), and stratum lacunosum-moleculare/dentate gyrus subgranular zone (SLM/SGZ) were grouped as neuropil-rich domains at the broad domain level. Domains mapping to white matter (WM.1, WM.2, and WM.3) were grouped as white matter domains at the broad domain level. Domains mapping to vascular and choroid plexus tissue were grouped as vascular/cerebrospinal fluid (CSF) domains at the broad domain level.

We found two domains each mapping to CA2-4 (CA2-4.1 and CA2-4.2) and CA1 (CA1.1 and CA1.2). Gene expression profiles for each of these domains were similar (**Extended Data Fig. 12D**). However, the number of nuclei per spot was higher in CA1.1 compared with CA1.2 and CA2-4.1 compared with CA2-4.2 (**Extended Data Fig. 12B**). Therefore, we collapsed CA2-4.1/CA2-4.2 to CA2-4 and CA1.1/CA1.2 to CA1 spatial domains for further analysis.

Spatial domain projection, Visium-SPG samples

We used the `RcppML` [32] R package to approximate `PRECAST` spatial domain membership for Visium-SPG samples, which were not included in the `PRECAST` clustering analysis. We used the `aggregateAcrossCells()` function from `scran` [109] to pseudobulk the logcounts matrix from Visium-H&E samples only, summing all logcounts for all genes across all spots in each domain. We pseudobulked the logcounts by domain such that our pseudobulked SPE object had 16 columns, one for each domain. We filtered lowly expressed genes using the `filterByExpr()` function from the `edgeR` package [117,118]. We rescaled columns of the pseudobulked matrix such that each column summed to 1, akin to the scaling diagonal used in NMF calculations in `RcppML` [32]. We then used the `project()` function from `RcppML` to estimate `PRECAST` spatial domain membership of each spot in the Visium-SPG samples. This function projects a linear factor model by solving this equation for *H* when given *A* and *W*:

$$A_{(i,j)} = W_{(i,k)} H_{(k,j)}$$

where *A* is a matrix with dimensions *i x j* and *W* and *H* are orthogonal, lower-rank matrices with dimensions *i x k* and *k x j*, respectively. In the context of `RcppML`, this function is generally used to project an NMF model of rank *k* trained in one dataset into a new dataset. We detail this process further in **Methods 4.8**. Here, we

instead used this function to predict `PRECAST` domain membership of each spot in the Visium-SPG samples (*H* matrix), given the Visium-SPG logcounts matrix (*A* matrix) and the pseudobulked domain-level matrix described above (*W* matrix). In this context, the *A* matrix was the Visium-SPG logcounts matrix, the *W* matrix was the pseudobulked logcounts matrix described above, and the *H* matrix was a matrix of spatial domain membership predictions for Visium-SPG spots. These predicted domain membership values were continuous. Therefore, for each spot, we assigned domain membership based on the highest predicted domain value.

Identification of thalamus and amygdala spots

We identified the presence of a small amount of thalamus tissue in 1 capture area using thalamus-specific marker *TCF7L2* (**Extended Data Fig. 14A-D**) [19,20]. Thalamus tissue clustered exclusively within SUB and SUB.RHP `PRECAST` domains. However, this capture area did not contain any subiculum or RHP tissue. The thalamic SUB/SUB.RHP spots were the only spots from SUB or SUB.RHP spatial domains in this capture area. These spots were not included in differential expression (DE) analyses.

We identified 3 total capture areas from two donors containing tissue from adjacent amygdala based on expression of several genes known to be enriched in amygdala, including *SLC17A6*, *CDH22* [21], *OPRM1* [23], and *CACNG4* [22] (**Extended Data Fig. 14E-J**). We also observed anatomy consistent with the transition between amygdala and HPC (**Extended Data Fig. 14E-F**). One of these 3 capture areas was entirely composed of amygdala tissue. These spots were not included in DE analyses. In the other 2 capture areas, amygdala tissue was clustered entirely within the RHP domain. These 2 capture areas did not include any other spots belonging to the RHP domain. These RHP spots were not included in DE analyses.

Pseudobulk processing

Following unsupervised clustering, Visium spots were pseudo-bulked by their spatial domain and capture area, as previously described [4]. Briefly, we summed the raw gene expression counts for across all spots in a given capture area in a given spatial domain. We performed this aggregation using the `aggregateAcrossCells()` function from `scran` [109]. We refer to these aggregated samples as pseudobulk samples. We discarded pseudobulk samples that were composed of less than 50 spots. We then identified and removed lowly expressed genes using the `filterByExpr()` function from the `edgeR` package [117,118]. Counts were normalized by library size, converted to counts per million (CPM), and log2 transformed using the `calcNormFactors()` function from `edgeR`. QC metrics were computed for each pseudobulk sample, again using `addPerCellQC()` from `scuttle` [110]. We applied a 2x MAD threshold to discard pseudobulk samples with very low detected genes.

For dimensionality reduction of pseudobulked data, we performed principal component analysis (PCA), retaining the first 100 principal components (PCs). We used the `getExplanatoryPCs()` function from the `scater` package [110] to compute, for each PC, the percent of variance explained by different metadata variables, including domain, donor, capture area, slide, age, sex, and QC metrics. We also used the `getExplanatoryVariables()` function from `scater` to compute, for each gene, the percent of variance explained by the same metadata variables.

Pseudobulk differential expression analysis

We performed differential expression (DE) analysis at both the domain and the broad domain levels using pseudobulk $\log_2$-transformed CPM values. We performed DE analysis to model differences in gene expression across domains or broad domains using `spatialLIBD` functions [108]. These functions are wrappers for various functions from the `limma` package [119]. We followed a modeling approach similar to Maynard et al. [4]. We accounted for age, sex, and slide as covariates. After model matrix construction, we ran `registration_block_cor()` from `spatialLIBD`, blocking by capture area. This function wraps

`duplicateCorrelation()` from `limma` to calculate correlation between pseudobulk samples within each block. This approach accounted for capture area-specific variation. We then fit three types of models as described in Maynard et al. [4], taking the block correlation into account. We refer to these models as ANOVA, enrichment, and pairwise models. For the ANOVA model, we used `registration_stats_anova()` from `spatialLIBD`, which wraps `lmfit()` and `ebayes()` from `limma` to test if expression of each gene is variable between domains. This model computed *F*-statistics for each gene across all domains. For the enrichment model, we used `registration_stats_enrichment()` from `spatialLIBD`, wrapping the same `limma` functions to test for gene expression differences between each domain and all other domains. For each gene and domain, this model computed fold change (FC), Student's *t*-test statistics, and two-tailed *p*-values. *P*-values were corrected for the false discovery rate (FDR) [120]. Genes were considered significant if $\log_2$FC > 1 or $\log_2$FC < -1 and FDR < 0.01. For the pairwise model, we used `registration_stats_pairwise()` from `spatialLIBD`, which wraps `lmfit()` and `ebayes()` from `limma` while also using `contrasts.fit()` from `limma` to introduce contrasts between each pair of domains. This model tested for gene expression differences between each pair of domains. For each gene and pair of domains, this model computed fold change (FC), Student's *t*-test statistics, and two-tailed *p*-values. *P*-values were corrected for FDR. Genes were considered significant if $\log_2$FC > 1 or $\log_2$FC < -1 and FDR < 0.01. For all models, we extracted $\log_2$FC, *F*-statistics or *t*-statistics, *P*-values, and FDR-corrected *P*-values.

Histologically-defined annotations

All histologically-defined annotations were performed by experienced neurobiologists (co-authors EDN, SCP). To annotate Visium-H&E samples, we built a temporary `shiny` application using `spatialLIBD::run_app()` [4,108]. We annotated the following anatomical regions: granule cell layer (GCL), subgranular zone (SGZ), molecular layer (ML), CA4 pyramidal cell layer (PCL-CA4), PCL-CA3, PCL-CA1, subiculum (SUB), stratum oriens (SO), stratum radiatum (SR), stratum lucidum (SL), stratum lacunosum-moleculare (SLM), white matter (WM), choroid plexus (CP), thalamus (THAL), and cortex (CTX) (**Extended Data Fig. 17**). We used marker genes from snRNA-seq studies in both human HPC [14,15] and mouse HPC [10,12]. We also used H&E images for histological reference. To identify GCL, we used *PROX1* and *SEMA5A*, but were also guided by high levels of hematoxylin staining due to densely packed granule cell bodies. To identify SGZ, we used *SEMA5A*, but not *PROX1*, expression. We also used *GAD1/GAD2* due to the presence of GABAergic neurons in SGZ [121]. To identify ML, we noted low UMIs and detected genes in areas between GCL and SLM. To identify PCL-CA4, we used *NECTIN3*, *AMPH*, *SEMA5A*, *SLC17A7*, and the presence of higher UMIs and detected genes compared with adjacent SGZ. For PCL-CA3, we used *NECTIN3*, *NECAB1*, *AMPH*, and *MPPED1*. For PCL-CA1, we used *FIBCD1*, *MPPED1*, *CLMP*, and *SLC17A7*. For SUB, we used *FN1*, *NTS*, and *SLC17A6*. For SO, we used *PVALB*, *SST*, *GAD1/GAD2*. For SR, we looked for regions between SLM and PCL-CA1/PCL-CA3 with high levels of *GFAP*, but lower levels of *MBP*/*MOBP* compared with SLM and WM. For SL, we looked for regions between SR and PCL-CA3 with low UMIs/detected genes and higher levels of *GFAP* and *MBP*/*MOBP* than adjacent PCL-CA3. For SLM, we noted a strip of darker staining compared with adjacent ML and SR along with expression of *MBP*/*MOBP* and *GFAP*. To identify WM regions, we looked for very high levels of oligodendrocyte markers such as *MBP*, *MOBP*, and *PLP1*. To identify CP, we used *TTR*, *PRLR*, and *MSX1*. CP tissue also had a distinctive sponge-like appearance, making visual identification easy. For thalamus, we used *TCF7L2*. Finally, for CTX, we used *CUX2*, *RORB* and *BCL11B*.

H&E Image Processing

The high-resolution images obtained as part of the 10x Genomics Visium protocol were processed using *VistoSeg*, a *MATLAB*-based pipeline that integrates gene expression data with histological data [122]. First, the

software splits the high-resolution H&E image of the entire slide into individual capture areas (.tif file format) and the capture areas of a single brain sample are rotated accordingly to orient to each other forming the hippocampal structure. The individual tif files are used as input to 1) `Loupe Browser` (10x Genomics) for fiducial frame alignment, and 2) `SpaceRanger` (version 1.3.0) to extract the spot metrics/coordinates. We then used the `VNS()` function from `VistoSeg` to obtain the initial nuclei segmentations. The `refineVNS()` function with adaptive thresholding was used to segment out the more accurate nuclei from background signals. Segmented regions within the size range of 40 to 4,000 pixels were only retained and a watershed was applied to split the nuclei clusters. These final segmentations were used to obtain the nuclei count per Visium spot. We took two approaches for spot level counting with different stringencies: 1) we counted all segmented nuclei whose centroids were within a given Visium spot and 2) we calculated the proportion of pixels within a given Visium spot in which a segmented nucleus was present. A function in *VistoSeg* then uses the spot metrics and coordinates from `SpaceRanger` to integrate the segmented nuclei counts with gene expression data.

Fluorescence Image Processing

We leveraged the immunofluorescence (IF) images in the Visium-SPG assay to directly quantify cell type abundance and provide an imaging-based calculation of cell counts for comparison with outputs from spot deconvolution tools. Starting from the `cyto` model provided by *Cellpose* 2.0, nuclear segmentations of Br3942 and Br8325 were iteratively improved in the *Cellpose* GUI, and a new model was trained until DAPI segmentations visually satisfied our accuracy standards. The final refined model was used to segment all 8 capture areas to produce masks. Segmentation masks were read into python 3.8.12, and mean fluorescence intensity for the NeuN, OLIG2, GFAP and TMEM119-marked channels was quantified using `skimage.measure.regionprops_table()` from the `scikit-image` [123] v0.19.2 library. For 2 capture areas 332_D1 and 335_D1 comprising of almost all HPC regions, an expert experimenter (co-author SCP) used `Samui Browser` [73] to randomly select and annotate a spatially diverse set of 200 cells of each immunolabeled cell type – e.g. neuron, oligodendrocyte, astrocyte, microglia, or other (i.e. presence of DAPI, but absence of other immunofluorescent signals). The resulting dataset included ~1000 cells with a cell type label, and the corresponding mean fluorescence intensities from the respective channels.

4.5 | snRNA-seq data generation

snRNA-seq data collection and sequencing

Using previously mentioned 100μm cryosections collected from each donor, we conducted single-nucleus RNA-sequencing (snRNA-seq) using 10x Genomics Chromium Single Cell Gene Expression V3 technology. Cryosections for each donor were first pooled with chilled Nuclei EZ Lysis Buffer (MilliporeSigma #NUC101) into a glass dounce. Sections were homogenized using 10-15 strokes with both pestles (loose and tight-fit). Homogenates were filtered through 70 μm mesh strainers and before centrifugation at 500 x g for 5 minutes at 4°C using a benchtop centrifuge. Nuclei were resuspended in fresh EZ lysis buffer, centrifuged again, and equilibrated in wash/resuspension buffer (1x PBS, 1% BSA, 0.2U/μL RNase Inhibitor). Nuclei were washed in wash/resuspension buffer and centrifuged 3 times. We labeled nuclei with Alexa Fluor 488-conjugated anti-NeuN (MilliporeSigma cat. #MAB377X), diluted 1:1000 in nuclei stain buffer (1x PBS, 3% BSA, 0.2U/μL RNase Inhibitor) by incubating at 4°C with continuous rotation. Following NeuN labeling, nuclei were washed once in stain buffer, centrifuged, and resuspended in wash/resuspension buffer. Nuclei were labeled with propidium iodide (PI) at 1:500 in wash/resuspension buffer before being filtered through a 35um cell strainer. We then performed fluorescent activated nuclear sorting (FANS) for using a Bio-Rad S3e Cell Sorter. Gating criteria were selected for whole, singlet nuclei (by forward/side scatter), G0/G1 nuclei (by PI fluorescence), and

neuronal nuclei (by Alexa Fluor 488 fluorescence). For nuclei from each donor, we first sorted 9,000 nuclei based on PI+ fluorescence, thereby including both neuronal and non-neuronal nuclei. Then, we sorted 9,000 additional nuclei into a separate tube based on both PI+ and NeuN+ fluorescence to facilitate enrichment of neurons. This approach initially gave N=20 for snRNA-seq (1 PI+ and 1 PI+NeuN+ sample for all 10 donors) with a total of 18,000 sorted nuclei per donor. Samples were collected over multiple rounds, each containing 1-3 donors for 2-6 samples per round. We had poor nuclei yields from one sequencing round, comprising both PI+ and PI+NeuN+ samples from 3 donors (**Extended Data Fig. 7**). We collected additional PI+ and NeuN+ samples from these 3 donors and performed nuclei sorting steps again, this time with better final nuclei yields. Therefore, we had a final N=26 for snRNA-seq (1 P1+ and 1 PI+NeuN+ sample for 7 donors, 2 PI+ and 2 PI+NeuN+ samples for 3 donors). All samples were sorted into reverse transcription reagents from the 10x Genomics Single Cell 3′ Reagents kit (without enzyme). Enzyme and water were added to bring the reaction to full volume. cDNA synthesis and subsequent sequencing library generation was performed according to the manufacturer's instructions for the Chromium Next GEM Single Cell 3' v3.1 (dual-index) kit (CG000315, revision E, 10x Genomics). Samples were sequenced on a Nova-seq (Illumina) at the Johns Hopkins University Single Cell and Transcriptomics Sequencing Core at a minimum read depth of 50000 reads per nucleus.

4.6 | snRNA-seq data analysis

<u>snRNA-seq data processing and quality control</u>

Following sequencing, we mapped reads to Genome Reference Consortium Human Build 38 (GRCh38) using `cellranger count` (version 7.0.0). Both exons and introns were included during alignment to increase library complexity. After alignment, we built a `SingleCellExperiment` [106] object with nucleus × gene UMI count matrices from `cellranger count`. Using the `emptyDrops()` function from the `dropletUtils` [124] Bioconductor package, we took a Monte Carlo simulation approach to identify and remove empty droplets and other sources of ambient RNA. We computed quality control metrics using the `addPerCellQC()` function from the `scuttle` Bioconductor package [110]. Neurons have larger library sizes and numbers of expressed genes compared with non-neuronal cells in human brain tissue [125]. PI+NeuN+ samples were expected to have higher levels of neurons than PI+ samples. Therefore, we performed outlier detection separately in PI+ and PI+NeuN+ samples. To identify poor-quality nuclei, we applied a 3x median absolute deviation (MAD) threshold for key QC metrics (percentage of UMIs mapping to mitochondrial genes, library size, and number of detected genes) [110]. These thresholds were computed on a per-sample basis. However, for 10 of the 26 samples, this automatic outlier detection method resulted in mitochondrial expression rate thresholds that were too high (9-27%) for effective quality control of snRNA-seq data. Six of these samples came from the same sample preparation round that resulted in low nuclei yields (**Extended Data Fig. 3**), indicating a potential batch-specific issue. For these 10 samples, we computed a shared median and MAD from the other 16 samples to estimate a more suitable mitochondrial expression rate threshold. Poor-quality nuclei were annotated and removed prior to downstream analyses. Following nuclei quality control, the dimensions of our dataset were 36,601 genes x 86,905 nuclei.

<u>snRNA-seq feature selection, dimensionality reduction, and clustering</u>

After this initial removal of nuclei, we noticed that one PI+NeuN+ sample had much more nuclei than expected (**Extended Data Fig. 8A-B**). For each sample, we once again examined the number of detected genes in retained nuclei and found that this sample had far more nuclei with low numbers of detected genes relative to other PI+NeuN+ samples. Many of these nuclei with low detected genes expressed *SYT1*, suggesting they might be low-quality neurons (**Extended Data Fig. 8C)**. We asked if these low-quality nuclei might cluster together and performed feature selection, dimensionality reduction, and clustering.

For feature selection, we used the `devianceFeatureSelection()` function from `scry` [126] to fit a Poisson model to gene counts across all cells and rank genes by deviance from this model. The top 2000 most deviant genes, or highly deviant genes (HDGs), were retained for dimensionality reduction and clustering. We calculated Pearson residuals from the Poisson model for these 2000 genes `using nullResiduals()` from `scry` [126]. With `runPCA()` from `scater` [110], we applied principal component analysis (PCA) to these Pearson residuals, retaining the top 100PCs. This approach approximates GLM-PCA, which avoids biases introduced by log-normalization of UMI count data [126]. To correct our dimensionality reduction results for batch effects, we followed PCA with mutual nearest neighbors (MNN) correction [127], correcting for sequencing round and donor-specific effects. For visualization purposes, we also produced a low-dimensional UMAP embedding [113]. We also normalized counts across nuclei. We first used the `quickCluster()` function from `scran` [109] to group nuclei with similar expression profiles. Then, we used the `calcSumFactors()` function from `scran` to deconvolve size factors for these clusters [128]. We normalized counts by these size factors and $\log_2$ transformed the normalized counts. Normalized counts were used for non-negative matrix factorization, but not clustering.

We used the `igraph` [129] and `scran` [109] packages to perform an initial round of Louvain clustering [130] on our PCA results using *k*=100 nearest neighbors and Jaccard weights. This resulted in 33 initial clusters (**Extended Data Fig. 8D**). We annotated these initial clusters as neuronal or non-neuronal based on expression of neuron-specific genes, such as *SYT1* (**Extended Data Fig. 8E**). Two neuronal clusters had much fewer detected genes relative to other neuronal clusters (**Extended Data Fig. 8F-G).** Approximately half of the nuclei in these two clusters were from the problematic PI+NeuN+ sample (**Extended Data Fig. 8I**). We reasoned that many of the nuclei in this sample were low-quality neurons. To prevent detrimental impact on downstream analyses, we took a conservative approach to this sample by removing all nuclei with fewer than 5000 nuclei in this sample only (**Extended Data Fig. 8H**). Following this removal step, the dimensions of our dataset were 36,601 genes x 80,695 nuclei.

We performed a second round of feature selection, dimensionality reduction, batch correction, and graph-based clustering using the same approach, this time using *k*=5 nearest neighbors in graph-based clustering to increase resolution. This resulted in 59 fine-grained clusters (**Extended Data Fig. 9A**). We annotated these second-round clusters as neuronal or non-neuronal based on expression of neuron-specific genes, such as *SYT1* (**Extended Data Fig. 9B-C**). We identified 3 low-quality neuron clusters based on low levels of detected genes (**Extended Data Fig. 9D**). We also identified 1 cluster of putative doublets, which were marked by coexpression of oligodendrocyte, astrocyte, OPC, and microglia markers in the same nuclei (**Extended Data Fig. 9E**). These clusters were composed of 4,027 total nuclei (~5% of all nuclei). We removed these 4 clusters (**Extended Data Fig. 9F-G**). Following this removal step, the dimensions of our dataset were 36,601 genes x 76,668 nuclei.

We performed a third round of feature selection, dimensionality reduction, batch correction, and graph-based clustering using the same approach, again using *k*=5 nearest neighbors in graph-based clustering. This resulted in 62 fine-grained clusters (**Extended Data Fig. 10A**). We annotated these third-round clusters as neuronal or non-neuronal based on expression of neuron-specific genes, such as *SYT1* (**Extended Data Fig. 10B-C**). We identified 1 small low-quality neuron cluster with low levels of detected genes relative to other neuron clusters (**Extended Data Fig. 10D**). We also identified 1 small cluster of putative doublets, which were marked by coexpression of oligodendrocyte, astrocyte, and OPC markers in the same nuclei (**Extended Data Fig. 10E**). We removed these 2 clusters (**Extended Data Fig. 10F**), leaving a total of 60 clusters for downstream analyses. Because these clusters were small (1,257 total nuclei, ~1.6% of all nuclei), we did not recluster the data after their removal (**Extended Data Fig. 10G**). Following this final removal step, the dimensions of our dataset were 36,601 genes x 75,411 nuclei.

<u>snRNA-seq cluster annotation</u>

We annotated clusters at 4 levels based on marker genes from previous snRNA-seq studies in both human [14,15] and mouse HPC [10–12]. In order of increasing resolution, we refer to these 4 levels as “broad” cell class labels (*k*=5), “mid” cell class labels (*k*=8), “fine” cell class labels (*k*=24), and “superfine” cell class labels (*k*=60), with each superfine cell class being one of the 60 final clusters (**Extended Data Fig. 29**). The “broad” cell class labels included neurons (excitatory and inhibitory, 52,000 nuclei total), astrocytes, oligodendrocytes, microglia, and other (comprised of OPCs, vascular cells, and CSF-related cells, 8,386 nuclei total). The “mid” cell class labels included excitatory neurons (41,253 nuclei), inhibitory neurons (10,747 nuclei), oligodendrocytes (6,787 nuclei), OPCs (1,662 nuclei), microglia and immune cells (3940 nuclei), astrocytes (4,298 nuclei), CSF-related cells (3,587 nuclei), and vascular cells (3,137 nuclei).

Non-pseudobulk differential expression

We used the `findMarkers()` function from the `scran` package to identify marker genes for mid, fine, and superfine cell class groupings using pairwise *t*-tests (**Supplementary Table 6, Supplementary Table 7, Supplementary Table 8**). We used `findMarkers()` to perform pairwise *t* tests between all pairs of clusters within each cell class grouping. We used parameters direction=”up” and pval.type=”all” to identify genes with a positive log-fold change and combining *p*-values across all pairwise comparisons for a given cluster within each cell class grouping.

Pseudobulk processing

Following unsupervised clustering, nuclei were pseudo-bulked by their superfine cell class and sample. We dropped all genes with zero counts across all nuclei and summed the raw gene expression counts for across all nuclei in a given sample in a given superfine cell class using the `aggregateAcrossCells()` function from `scran` [109]. We discarded pseudobulk samples that were composed of less than 50 nuclei. We then identified and removed lowly expressed genes using the `filterByExpr()` function from the `edgeR` package [117,118]. Counts were normalized by library size, converted to counts per million (CPM), and $\log_2$ transformed using the `calcNormFactors()` function from `edgeR`. QC metrics were computed for each pseudobulk sample, again using `addPerCellQC()` from `scuttle` [110]. We applied a 3x MAD threshold to discard pseudobulk samples with very low detected genes.

Pseudobulk differential expression analysis

We performed differential expression (DE) analysis at the mid, fine, and superfine cell class levels (**Supplementary Table 9, Supplementary Table 10, Supplementary Table 11)** using pseudobulk $\log_2$-transformed CPM values. We again used `spatialLIBD` functions [108] for DE. We accounted for age, sex, donor, and sequencing round as covariates. After model matrix construction, we ran `registration_block_cor()` blocking by sample to account for sample-specific variation. We then fit ANOVA, enrichment, and pairwise models as described in Maynard et al. [4], taking the block correlation into account. For the ANOVA model, we used `registration_stats_anova()` to test for gene-level variance between clusters in each cell class grouping. This model computed *F*-statistics for each gene across all clusters in each cell class grouping. For the enrichment model, we used `registration_stats_enrichment()` to test for gene expression differences between each cluster and all other clusters within each cell class grouping. For each gene and cluster, this model computed fold change (FC), Student’s *t*-test statistics, and two-tailed *p*-values. *P*-values were corrected for the false discovery rate (FDR) [120]. Genes were considered significant if $\log_2$FC > 1 or $\log_2$FC < -1 and FDR < 0.01. For the pairwise model, we used `registration_stats_pairwise()` to test for gene expression differences between each pair of clusters within each cell class grouping. For each gene and pair of clusters, this model computed fold change (FC), Student’s *t*-test statistics, and two-tailed *p*-values. *P*-values were corrected for FDR. Genes were

considered significant if $\log_2FC > 1$ or $\log_2FC < -1$ and FDR < 0.01. For all models, we extracted $\log_2FC$, *F*-statistics or *t*-statistics, *p*-values, and FDR-corrected *p*-values.

4.7 | Spot-level deconvolution

Spot deconvolution benchmarking

Given that individual Visium spots often contain multiple cell types, we performed spot deconvolution to better understand the cellular composition of spots mapping to unsupervised spatial domains. While several algorithms have been developed to predict cell type proportions within individual Visium spots using single cell reference data, they have not yet been comprehensively benchmarked across various brain regions due to limited availability of complete reference datasets which contain 1) paired SRT and snRNA-seq data from the same donors, 2) known cell type abundances in each gene expression spot, 3) anatomical regions enriched for specific cell types. Therefore, we generated the first gold standard spot deconvolution dataset for 4 broad cell types in postmortem human anterior HPC using the Visium Spatial Proteogenomics assay (Visium-SPG), which replaces H&E histology with immunofluorescence staining to label proteins of interest with fluorescent dyes (**Fig. 1A,D)** similar to previous work in the DLPFC **[5]**. We selected 2 out of the 10 brain samples and performed immunofluorescent staining for established cell type-specific proteins, including NEUN (marking neurons), OLIG2 (marking oligodendrocytes), GFAP (marking astrocytes), and TMEM119 (marking microglia) (**Fig. 1D**, **Extended Data Fig. 1B**). Following multispectral fluorescence imaging, we proceeded with the standard Visium protocol to generate corresponding gene expression libraries from the same tissue section.

Marker gene detection

We identified a set of 25 robustly expressed marker genes for snRNA-seq clusters at both fine (**Fig. 3A**) and mid resolutions (**Extended Data Fig. 29**). The get_mean_ratio2() function from `DeconvoBuddies` (version 0.99.0) [30] was applied to rank each gene in the snRNA-seq data in terms of its suitability as a marker for each possible target cell type at both resolutions. This method, termed “mean ratio”, determines the ratio of expression between a target cell type and the next-highest-expressing cell type (**Extended Data Fig. 33**). Mitochondrial genes and genes not present in the spatial data were excluded from the ranking process. Three snRNA-seq clusters from the fine grouping (“HATA”, “Amy” and “GABA.PENK") were also excluded as these weren't represented in any Visium-SPG samples, clearly mapped to amygdala, and could occlude marker gene identification for other pyramidal neuron clusters (**Extended Data Fig. 28C-D**). We retained the “thalamus” cluster because one of the Visium-SPG samples included a very small amount of thalamus (**Extended Data Fig. 1B, Extended Data Fig. 14**) which was transcriptionally distinguishable from hippocampal cells. The top-25-ranked markers by mean ratio were taken for each cell type, and, to ensure greater expression in the target cell type than all others, we verified that the mean ratio exceeded 1 for each gene. This resulted in a total of 200 mid-resolution markers and 525 fine-resolution markers used downstream for spot deconvolution (**Supplementary Table 12**). To confirm the utility of the snRNA-seq-derived marker genes for spot deconvolution, we evaluated their proportion of nonzero expression in Visium-SPG data (**Extended Data Fig. 34**) and confirmed that 25 genes per cell type were sufficient to identify the expected cell type distribution patterns (**Fig. 4B(ii), Extended Data Fig. 35**). Using Granule Cell (GC) excitatory neurons as an example, we found consistent spatial localization between Granule Cell Layer (GCL) marker gene *PPFIA2* and the top 25 GC marker genes identified in snRNA-seq data (**Fig. 4A, Extended Data Fig. 34**).

Applying spot deconvolution softwares

Having verified the selection of robust marker genes for each cell type at both fine and mid resolutions, we ran `cell2location` [27], `RCTD` [28] and `Tangram` [29] and calculated the predicted cell type counts and proportions per spot. `Cell2location` v0.8a0, `RCTD` v2.2.0 and `Tangram` v1.0.2 were applied with default

parameters following the respective tutorials provided by their authors, with a few exceptions. The marker genes derived from the `Deconvobuddies` package were used as training genes. This strategy was effective for `cell2location` and `Tangram`. However, in `RCTD` it failed to predict cell types in spots with fewer UMIs (~<100) after filtering the genes to include only the user derived marker genes (**Extended Data Fig. 36**). We observed that `RCTD` dropped a high number of spots when user-defined marker genes were used (**Extended Data Fig. 36A**), but performance improved when its default marker gene detection strategy was employed (**Extended Data Fig. 36B**). `Tangram` requires an accurate cell count per spot, and thus was provided total cell counts per spot derived through nuclei segmentation with `Cellpose` v2.0 (Visium-SPG images) and through `VistoSeg` (H&E images). `Cell2location` and `RCTD` require an average count of nuclei per spot, which was calculated as 5 across all 8 capture areas. While `Tangram` and `Cell2location` predict cell type abundances in each spot, `RCTD` predicts the cell type proportion per spot. Therefore, to more directly compare across algorithms, we multiplied the cell type proportions obtained from `RCTD` by the nuclei counts to obtain abundances for each spot.

<u>Evaluating performance of spot deconvolution methods</u>

First, using the IF images, we calculated the actual number of neurons, astrocytes, microglia, and oligodendrocytes per spot by segmenting individual nuclei and implementing a classification and regression tree (CART) approach in `scikit-learn` [131] to categorize nuclei into either one of the 4 immunolabeled cell types or “other” (**Fig. 4B(ii)**, **Extended Data Fig. 37**). This provided an orthogonal metric to classify cell type identity within each spot. The cell type counts estimated by each algorithm in the Visium-SPG data at fine and mid resolutions were “collapsed” down to the broad categories (neuron, astrocytes, microglia, oligodendrocytes, and other (comprised of OPCs, vasculature, and CSF-related cells)) to enable direct comparison with the CART-calculated counts. We then visually compared the predicted cell type (of mid cell class collapsed down to broad cell class) proportion per spot from the 3 softwares to the CART-estimated proportions of immunolabeled cells (**Extended Data Fig. 38**). The collapsed counts from mid and fine resolutions were compared against each other which should theoretically match (**Fig. 4C**). To quantify the comparisons, the “other” cell type was excluded from the benchmarking analyses and counts for each software tool and CART-quantified cell type were summed across each Visium-SPG section. Totals for each software method (gene expression) were compared against the CART-predicted (immunofluorescence) cell count totals using Pearson correlation and root mean squared error (RMSE) at mid (**Extended Data Fig. 39A**) and fine (**Extended Data Fig. 39B**) levels (**Supplementary Table 13**). Counts for each software tool and `CART` predictions were normalized to add to 1 across the four cell types which allowed the calculation of the Kulback-Lieber divergence (KL divergence) from each software tool’s predictions to the `CART` predictions. This treats the CART-predicted cell type composition as a ground-truth probability distribution that each software tool is attempting to estimate (**Extended Data Fig. 39A**). Given the challenge of deconvolving transcriptionally fine cell types, our Visium-SPG benchmark analyses confirmed `RCTD` showed the most consistent performance at both mid and fine resolution across all cell types and samples (**Fig. 4B-D(i)**, **Extended Data Fig. 39**).

Having evaluated the performance on the Visium-SPG data, we ran all 3 methods on the H&E data for which orthogonal measures of spot-level cellular composition were not available. For each cell type, we calculated the predicted cell type proportion per spot across all spatial domains determined by `PRECAST` (**Fig. 2**). Across all cell types, `RCTD` and `Cell2location` showed similar domain-mapping performance; however, `cell2location` made more conservative predictions compared to `RCTD`, while `RCTD` was able to accurately capture the dynamic changes in cell type proportions (**Fig. 4D(ii)**). We observed that `Tangram` showed consistency between mid and fine level predictions (**Extended Data Fig. 41**) but lacked spatial

specificity with more false positive neuronal cell types across all spatial domains, whereas `RCTD` predictions were more in line with expectations.

4.8 | Non-negative matrix factorization (NMF)

NMF factorization

We used non-negative matrix factorization (NMF) to identify continuous patterns of gene expression within our snRNA-seq data. NMF is a dimensionality reduction technique for pattern recognition [132] used across various disciplines, including transcriptomics [31,34]. NMF decomposes a matrix *A* into two lower-rank, orthogonal matrices *w* and *h*:

$$A_{(i,j)} \approx W_{(i,k)} H_{(k,j)}$$

where *k* is rank of the factorization and both *w* and *h* are constrained to have only non-zero elements. The decomposition is performed iteratively until the difference between *A* and the product of *w* and *h.* Iterations continue until the change in mean squared error (MSE) between successive iterations falls below a specified level (tolerance of fit). We used the `RcppML` package for NMF [32], which also features a scaling diagonal *D* such that:

$$A \approx WDH$$

where *D* diagonalizes columns in *W* and rows in *H* to sum to 1. In our case, *k*=100 patterns and *A* was the normalized snRNA-seq counts matrix with dimensions 36,601 genes x 75,411 nuclei. Therefore, *w* had dimensions 36,601 genes x 100 patterns and *h* had dimensions 100 patterns x 75,411 nuclei.

We used the `singlet` and `RcppML` packages for NMF analysis. These packages leverage the `RcppML` implementation of NMF for use with snRNA-seq/scRNA-seq data. `RcppML` is much faster and more memory efficient than other NMF methods [32]. Additionally, its diagonal scaling approach allows for L1/L2 regularization and reproducible NMF loadings.

We used the `cross_validate_nmf()` function from `singlet` to perform cross-validation of different NMF ranks [133]. For a given number of replicate runs, this function randomly splits the data into training and test sets of a user-specified size. For each replicate, NMF is performed for a given number of ranks in the training set until the tolerance threshold is met. Model overfitting is also measured by a tolerance of overfit metric. After 5 iterations, the factorization's performance is evaluated in the test set and test set MSE is calculated. At each successive iteration thereafter, the test set MSE is calculated compared to the test set MSE at 5 iterations. If the test set MSE for a given iteration is greater than test set MSE at iteration 5 by a specified value, the model is considered to be overfit. We cross-validated using 3 replicates with parameters tolerance of fit=10^{-3} (`singlet`-recommended threshold for cross-validation), tolerance of overfit=10^{-4}, and L1=0.1 across ranks *k*=5, 10, 50, 100, 125, 150, and 200. Test set MSE at tolerance of fit decreased dramatically until k=100, and did not clearly decrease at *k*=125, *k*=150, or *k*=200 (**Extended Data Fig. 42**). Additionally, model overfitting was observed at *k*=150 and *k*=200 for all 3 replicates. Because models with *k*=100 and k=125 performed similarly and did not overfit the test data, we *k*=100 for the sake of interpretability, reasoning that 100 patterns would be more easily annotated than 125. We performed the final factorization using k=100 and tolerance of fit=10^{-6} (`singlet`-recommended threshold for high-quality factorization results >= 10^{-5}).

NMF annotation, marker gene detection, and functional enrichment analysis

We first annotated patterns mapping to *a priori* identified cell classes. We aggregated NMF weights for all nuclei within each class. We identified 70 patterns mapping to 55 of the 60 superfine cell classes (**Extended Data Fig. 43**). Because NMF patterns were normalized to sum to 1, aggregating pattern weights across specific cell class categories gave us the proportion of each pattern mapping to each cell class.

We annotated patterns mapping to batch effects and other sample-level variables. Using the same one-hot encoding and correlation strategy, we correlated the transposed *H* matrix with various sample-level variables, including donor, sample, sequencing round, sex, and age. We also correlated with number of detected genes, library size, and mitochondrial expression rate to identify patterns mapping to quality control variables. We identified 7 patterns mapping to sample-level or technical variables. These included 2 patterns mapping mostly to specific donors, 2 patterns mapping to sex (1 to males, 1 to females, mostly marked by Y and X linked genes), 1 pattern mapping to mitochondrial reads, and 1 pattern correlated with library size and number of detected genes (**Extended Data Fig. 45**). We also identified 1 pattern that appeared to capture variance related to choroid plexus "smearing" and contamination, a known phenomenon in brain transcriptomics studies (**Extended Data Fig. 46**) [35]. These patterns were excluded from downstream analyses.

We annotated patterns mapping to biological processes using functional enrichment analysis. After removing patterns mapping to batch effects and sample-level variables, we removed from the *W* matrix genes with zero weights across all patterns, non-protein coding genes, and mitochondrial genes. We identified sets of marker genes for each remaining pattern using `patternMarkers()` [36] from the `CoGAPS` [134] package. Using these gene sets, we used `clusterProfiler` [135] to perform over-representation analysis for GO terms across all ontologies: biological process (BP), cellular component (CC), and molecular function (MF) (**Extended Data Fig. 47**)[37]. We classified non-cell type-specific patterns based on both `patternMarkers()` output and over-representation analysis results.

<u>NMF pattern transfer</u>

We took a transfer learning approach to predict pattern weights for the NMF patterns identified in human snRNA-seq data across other datasets. We used the `project()` function from the `RcppML [32]` package for this purpose. This function predicts sample-level weights in a target dataset for NMF patterns defined in a training dataset by solving the following equation for *H'* when given *A'* and *W*:

$$A'_{(i',j')} = W_{(i',k)} H'_{(k,j')}$$

In this context, *A'* is a target dataset with dimensions *i'* genes x *j'* samples, *W* is the original gene-level NMF weights matrix defined using human HPC snRNA-seq data, and *H'* is the predicted sample-level weight matrix for the target dataset. This function makes use of a fast implementation of non-negative least squares [136].

We used this approach to project NMF patterns defined in human HPC snRNA-seq data into three different datasets: human HPC SRT data, epi-retro-seq data in a mouse model linking single-nucleus methylation sequencing in neurons with their axon projection targets [43], and snRNA-seq data in a mouse model of induced electroconvulsive seizures (ECS) [44]. For pattern transfer to human HPC SRT data, we used library size-normalized, $\log_2$ transformed counts as the *A'* matrix. For pattern transfer to epi-retro-seq data, we first subset the dataset to HPC and RHP neurons only. We processed the data as described in the original study [43]). We used log-scaled, negated methylation counts given that cytosine methylation has an inverse relationship with gene expression. We identified human orthologs for mouse genes in this dataset and removed all genes without orthologs. We matched these orthologs with genes in the *W* matrix and discarded any genes not included in both datasets. We used this log-scaled, negated methylation counts matrix as the *A'* matrix. For pattern transfer to mouse ECS snRNA-seq data, we again identified human orthologs for mouse genes in this dataset and removed all genes without orthologs. We matched these orthologs with genes in the *W* matrix and discarded any genes not included in both datasets. We used normalized, $\log_2$ transformed counts from the mouse ECS snRNA-seq dataset as the *A'* matrix. For the sake of interpretability, patterns were again normalized to sum to 1 in each target dataset following pattern transfer.

4.9 | Stratified Linkage disequilibrium score regression (S-LDSC)

Prior to S-LDSC, we defined gene sets for each set of labels. For snRNA-seq superfine cell classes and SRT spatial domains, we used *t*-statistics from our pseudobulk DE analysis enrichment model (each domain/superfine cell class vs all other domains/cell classes). For each unique domain or superfine cell class, genes were ranked by enrichment model *t*-statistics. The top 10% of genes by t-statistic were selected as the gene set for each unique domain or superfine cell class. For NMF patterns, we first calculated *z*-scores for each gene across all patterns not mapping to batch effects or technical variables. We then ranked genes by *z*-score for each pattern. The top 10% of genes by *z*-score were selected as the gene set for each unique NMF pattern. To construct a genome annotation for each unique label (spatial domain, snRNA-seq superfine cell class, or NMF pattern), we added a 100Kb window upstream and downstream of the transcribed region of each gene in that label's corresponding gene set.

We performed stratified LD score regression (S-LDSC) to evaluate the enrichment of heritability of brain-related traits for gene sets defined for different domains. We also included two non-brain traits, human height and type 2 diabetes, as negative controls to examine whether our findings are specific to brain-related traits. We downloaded GWAS summary statistics of each trait [53–68]. Following recommendations from the LDSC resource website (https://alkesgroup.broadinstitute.org/LDSCORE), S-LDSC was run for each gene set with the baseline LD model v2.2 that included 97 annotations to control for the LD between variants with other functional annotations in the genome. We used HapMap Project Phase 3 SNPs as regression SNPs, and 1000 Genomes SNPs of European ancestry samples as reference SNPs, which were all downloaded from the LDSC resource website. To evaluate the unique contribution of gene sets to trait heritability, we utilized the metric from S-LDSC: the *z*-score of per-SNP heritability. This metric allows us to discern the unique contributions of candidate annotations while accounting for contributions from other functional annotations in the baseline model. The *p*-values are derived from the *z*-score assuming a normal distribution and FDR was computed from the *p*-values based on Benjamini & Hochberg procedure.

4.10 | Data Visualization

<u>Multi-array integration for Visium-SPG sample visualization</u>

To enhance interactive visualization incorporating IF images alongside transcriptomic and spatial domain clustering, we utilized a robust, web-based data visualization tool tailored for spatially-resolved transcriptomics data, `Samui Browser` [73]. We integrated all capture areas from each donor to enable visualization of the intact hippocampal structure by performing the following steps, based on a previously described strategy [137]. First, we imported the high-resolution single-channel (DAPI) images from the `SpaceRanger` (version 2.0.0) outputs in `ImageJ` [138] to derive approximate transformations for all capture areas per donor. Due to image size constraints in `ImageJ`, high-resolution images featuring visible tissue landmarks were employed instead of full-resolution images or low resolution counterparts which lacked sufficient landmarks for accurate alignment. The "Tracing and Evaluation of Neural Anatomy using MipMaps" (`TrakEM2`) plugin [138–140] facilitated capture area alignment to reconstruct the hippocampal structure. Final transformations were exported in XML format. These transformations were extracted from the XML file of all capture areas per donor, then scaled to full resolution by applying scaling factors from the spaceranger JSON file using custom Python (v3.10.12) code, in particular `PIL` v10.0.0 and `numpy` v1.24.4. Then, the scaled transformations were applied to the full-resolution capture areas for each channel, and the transformed capture areas were placed onto a blank image encompassing the entire hippocampal structure from all capture areas from each donor. This composite image served as the basis for visualization in the `Samui Browser`. We then applied the same transformations to the Visium spot coordinates, where spot coordinates from all capture areas for each donor

were concatenated into a single file. Overlapping spots from capture areas with wrinkled or curled tissue sections were disregarded. A `SpatialExperiment` object per donor was converted into an `AnnData` object in R, ensuring compatibility with the `Samui Browser`. Specifically, the R function `zellkonverter::SCE2AnnData()` is used to convert the existing `SpatialExperiment` into an `AnnData` object, which allows gene expression and phenotype data to be accessible within python and therefore usable with the `Samui` API. We then used the combined image, spot coordinates and AnnData object to create the final `Samui` directory for each donor.

Multi-array integration for Visium-H&E sample visualization

A similar approach was also used for the Visium-H&E samples. The `Samui Browser` [30] was used as the final visualization tool. Using the `Fiji` distribution of `ImageJ`, images of the individual Visium slide capture areas were loaded into `ImageJ` and manually aligned. The transformations were saved as XML files. The `Spaceranger` JSON files were used to scale transformations to full resolution. The `zellkonverter::SCE2AnnData()` function was used to convert the existing `SpatialExperiment` into an `AnnData` object for use by `Samui Browser`. As in the Visium-SPG example, above, we used the combined image, spot coordinates and AnnData object to create the final product that is ready to be imported into `Samui`. All Visium data is available in a joint `Samui` browser.

iSEE apps

To enable exploration of the data, we created *iSEE* websites for pseudobulked snRNAseq and pseudobulked spatial data as previously described [5,72]. All interactive websites are available at research.libd.org/spatial_hpc/.

4.11 | Data and Code Availability

Raw and processed data are available through Gene Expression Omnibus (GEO) under accession GSE264624 [74]. The code for this project is publicly available through GitHub at https://github.com/LieberInstitute/spatial_hpc. Analyses were performed using R version 4.3.2 with Bioconductor version 3.17 unless otherwise noted. Image alignment and transformation was performed with `ImageJ` (version 2.14.0).

4.12 | Acknowledgements, Funding, Authorship Contributions

Acknowledgements:

Portions of some figures were created with BioRender.com. We thank the LIBD neuropathology team, particularly James Tooke and Amy Deep-Soboslay, for curation of the brain samples and assistance with tissue dissections. We thank the staff and physicians at the brain donation sites, and the generosity of the brain donors and their families, without whom this work would not be possible. We thank Daniel R Weinberger, members of the Hansen-Hicks working group and the Spatial LIBD team for feedback on the manuscript. Finally, we thank the families of Connie and Steve Lieber and Milton and Tamar Maltz for their generous support.

Funding:

This project was supported by U01MH122849 (KM), R21AG083328 (SCP/SCH), R01MH105592 (KM), and the Lieber Institute for Brain Development.

Conflict of Interest:
The authors have no declared conflicts of interests.

Author Contributions:
Conceptualization: KM, SCP, SCH
Methodology: EDN, MT, JY, SCH
Software: EDN, MT, ADR, RAM, NJE, LCT, JY, SH, SCH
Validation: EDN, ADR, SCP
Formal Analysis: EDN, MT, ADR, JY, SCH
Investigation: EDN, ADR, HRD, EAP, SHK, SVB, UMK, SCP
Resources: JEK, TMH
Data Curation: EDN, MT, ADR, SCH
Writing – original draft: EDN, MT, KRM, KM, SCP, SCH
Writing – review & editing: EDN, MT, ADR, HRD, RAM, SVB, SH, KRM, KM, SCP, SCH
Visualization: EDN, MT, ADR
Supervision: KRM, KM, SCP, SCH
Funding acquisition: KRM, KM, SCP, SCH
Project Administration: KRM, KM, SCP, SCH

Author ORCID:
- Erik D. Nelson (https://orcid.org/0000-0001-8477-0982)
- Madhavi Tippani (https://orcid.org/0000-0002-6465-6418)
- Anthony D. Ramnauth (https://orcid.org/0000-0001-5373-0857)
- Heena R. Divecha (https://orcid.org/0000-0002-1959-0675)
- Ryan A. Miller (https://orcid.org/0000-0003-3477-7443)
- Nicholas J. Eagles (https://orcid.org/0000-0002-9808-5254)
- Elizabeth A. Pattie (https://orcid.org/0009-0009-5729-4971)
- Sang Ho Kwon (https://orcid.org/0000-0001-5328-0956)
- Svitlana V. Bach (https://orcid.org/0000-0003-1134-9585)
- Uma M. Kaipa (https://orcid.org/0000-0002-4931-1753)
- Jianing Yao (https://orcid.org/0009-0004-8213-2118)
- Joel E. Kleinman (https://orcid.org/0000-0002-4210-6052)
- Leonardo Collado-Torres (https://orcid.org/0000-0003-2140-308X)
- Shizhong Han (https://orcid.org/0000-0002-5114-6742)
- Kristen R. Maynard (https://orcid.org/0000-0003-0031-8468)
- Thomas M. Hyde (https://orcid.org/0000-0002-8746-3037)
- Keri Martinowich (https://orcid.org/0000-0002-5237-0789)
- Stephanie C. Page (https://orcid.org/0000-0002-1951-7398)
- Stephanie C. Hicks (https://orcid.org/0000-0002-7858-0231)

Figures

Fig. 1. Experimental design to generate paired single-nucleus RNA-sequencing (snRNA-seq) and spatially-resolved transcriptomics (SRT) data in the human hippocampus.

(**A**) Postmortem human tissue blocks containing the anterior hippocampus were dissected from 10 adult neurotypical brain donors. Tissue blocks were scored and cryosectioned for snRNA-seq assays (red), and placement on Visium slides (Visium-H&E, black; Visium Spatial Proteogenomics (Visium-SPG), grey).

(**B**) 10µm tissue sections from all ten donors were placed onto 2-5 capture areas to include the extent of the HPC (*n*=36 total capture areas), for measurement with the 10x Genomics Visium-H&E platform.

(**C**) 10µm tissue sections from two donors were placed onto 4 capture areas (*n*=8 total capture areas) for measurement with the 10x Genomics Visium-SPG platform.

(**D**) Tissue sections (2-4 100µm cryosections per donor) from all ten donors were collected from the same tissue blocks for measurement with the 10x Genomics 3' gene expression platform. For each donor, we sorted on both PI+ and NeuN+, and replicate samples were collected from three donors for a total of *n*=26 total snRNA-seq libraries.

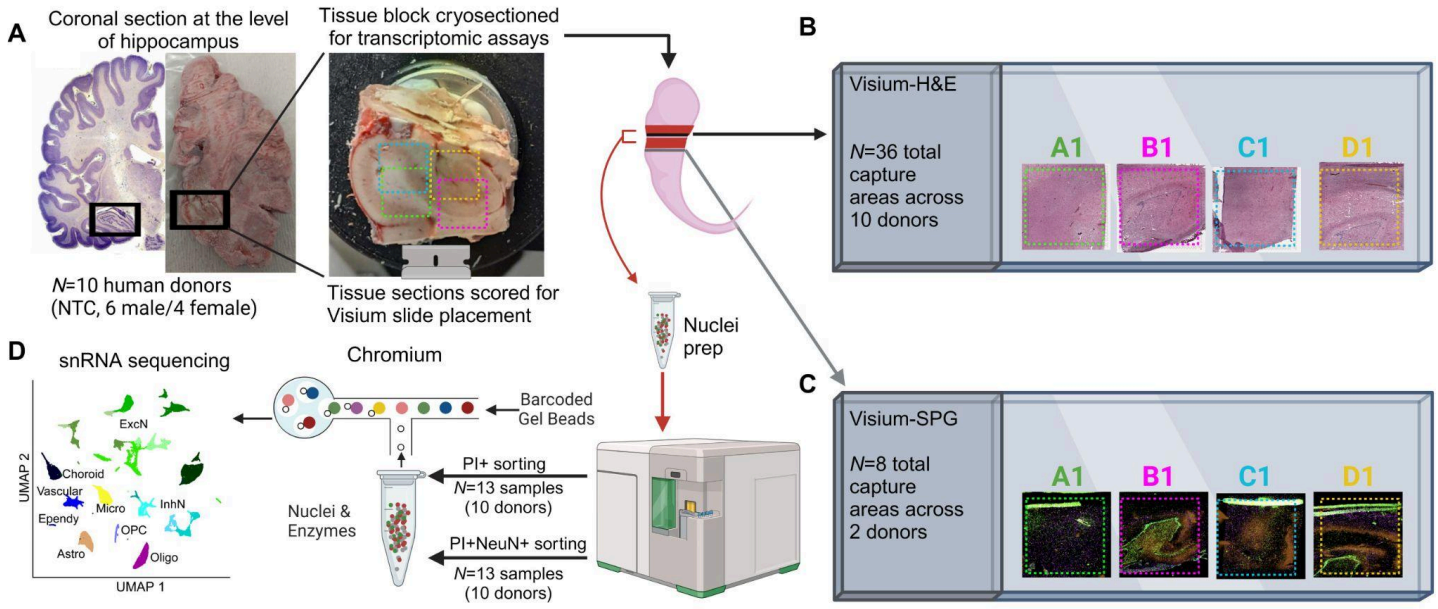


Fig. 2. Spatial domain detection and differential expression (DE) in the human hippocampus using spatially-resolved transcriptomics data.

(**A**) Integrated and merged spot plot of four Visium capture areas from the same donor (Br3942) with spots colored by annotated data-driven spatial domains (*k*=18) using `PRECAST`. CA1.1/CA1.2 were collapsed to CA1 and CA2-4.1/CA2-4.2 were collapsed to CA2-4. See **Supplementary Table 2** for abbreviations.

(**B**) Schematic illustrating pseudobulking approach, which collapses the spot-level to spatial domain-level data within each capture area.

(**C**) First two principal components (PCs) of the pseudobulked samples with colors corresponding to spatial domains and labeled ovals representing four broad classes (neuron cell body-enriched, neuropil-enriched, white matter-enriched, vasculature-enriched)

(**D**) Heatmap showing marker gene expression (rows) across annotated `PRECAST` spatial domains (columns). For visualization purposes, domains annotated as CA2-4.1 and CA2-4.2 were collapsed as CA2-4, and domains annotated as CA1.1 and CA1.2 were collapsed as CA1. Annotation across the top shows broad class grouping.

(**E**) Spot plot showing spatial expression of *PPFIA2*, a marker for the GCL. Spots are filled by $\log_2$ normalized counts. Spot borders are colored by broad domain.

(**F)** Spot plot showing expression of *PRKCG*, which is enriched in CA1-4 domains. Spots are filled by $\log_2$ normalized counts. Spot borders are colored by broad domain.

(**G**) Spot plot showing expression of *APOC1*, which is enriched in ML, SL-SR, and SR-SLM domains. Spots are filled by $\log_2$ normalized counts. Spot borders are colored by broad domain.

(**H**) Spot plot showing expression of *SFRP2*, a marker for the SLM.SGZ domain. Spots are filled by $\log_2$ normalized counts. Spot borders are colored by broad domain.

(**I**) Volcano plots illustrating results from DE analysis for each broad-level domains, with log2 fold change on the x-axis and FDR adjusted, $-\log_{10}$ transformed *p*-values on the y-axis. Genes colored blue pass the FDR threshold (FDR adjusted *p*-value < 0.01), while genes colored red pass both FDR and log2 fold change thresholds (FDR adjusted *p*-value < 0.01 and log2 fold change > 1). Top DE genes for each broad domain grouping are labeled.

(**J**) Boxplots showing expression of marker genes for neuron-enriched regions (*CLSTN3*), neuropil-enriched regions (*SLC1A3*), white matter regions (*SHTN1*), and vascular/CSF regions (*TPM2*). Each data point represents a pseudobulked sample. PRECAST clusters are on the x-axis and counts per million (cpm) for each gene are on the y-axis. Boxes are colored by broad cluster grouping.

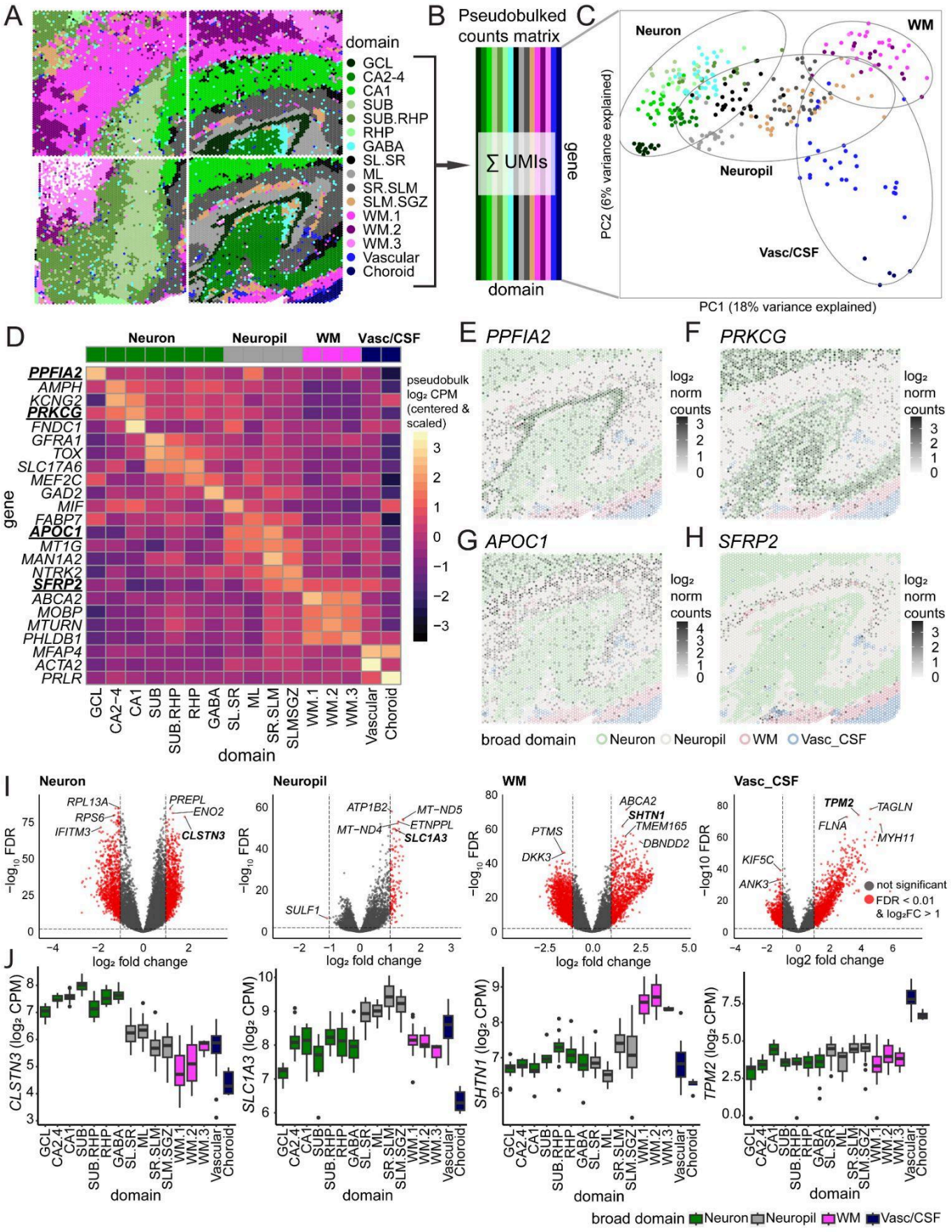

A
domain
GCL
CA2-4
CA1
SUB
SUB.RHP
RHP
GABA
SL.SR
ML
SR.SLM
SLM.SGZ
WM.1
WM.2
WM.3
Vascular
Choroid
B
Pseudobulked counts matrix
∑ UMIs
gene
domain
C
Neuron
WM
Neuropil
Vasc/CSF
PC2 (6% variance explained)
PC1 (18% variance explained)
D
Neuron
Neuropil
WM
Vasc/CSF
PPFIA2
AMPH
KCNG2
PRKCG
FNDC1
GFRA1
TOX
SLC17A6
MEF2C
GAD2
MIF
FABP7
APOC1
MT1G
MAN1A2
NTRK2
SFRP2
ABCA2
MOBP
MTURN
PHLDB1
MFAP4
ACTA2
PRLR
gene
GCL
CA2-4
CA1
SUB
SUB.RHP
RHP
GABA
SL.SR
ML
SR.SLM
SLMSGZ
WM.1
WM.2
WM.3
Vascular
Choroid
domain
pseudobulk log₂ CPM (centered & scaled)
E PPFIA2
F PRKCG
G APOC1
H SFRP2
log₂ norm counts
broad domain
Neuron
Neuropil
WM
Vasc_CSF
I
Neuron
Neuropil
WM
Vasc_CSF
RPL13A
RPS6
IFITM3
PREPL
ENO2
CLSTN3
ATP1B2
MT-ND5
MT-ND4
ETNPPL
SLC1A3
SULF1
PTMS
DKK3
ABCA2
SHTN1
TMEM165
DBNDD2
TPM2
TAGLN
FLNA
MYH11
KIF5C
ANK3
−log10 FDR
log₂ fold change
log2 fold change
not significant
FDR < 0.01 & log₂FC > 1
J
CLSTN3 (log₂ CPM)
SLC1A3 (log₂ CPM)
SHTN1 (log₂ CPM)
TPM2 (log₂ CPM)
domain
broad domain
Neuron
Neuropil
WM
Vasc/CSF

Fig. 3. **Cell type identification and differential expression (DE) in the human hippocampus using single-nucleus RNA-sequencing.**

(**A**) UMAP representation of the snRNA-seq dataset. Points are colored and labeled by fine cell class. See **Supplementary Table 2** for abbreviations.

(**B**) Stacked bar plots showing proportions of nuclei in each fine cell class by donor, by sort strategy (PI+ or PI+NeuN+), and across the overall dataset.

(**C**) Violin plots showing differential expression of marker genes across cell classes.

(**D**) Heatmap showing differential expression of marker genes across fine cell classes, with rows as genes and columns as superfine cell classes. Columns are annotated by cell class and broad cell class. Heatmap is colored by mean $\log_2$ transformed, normalized counts.

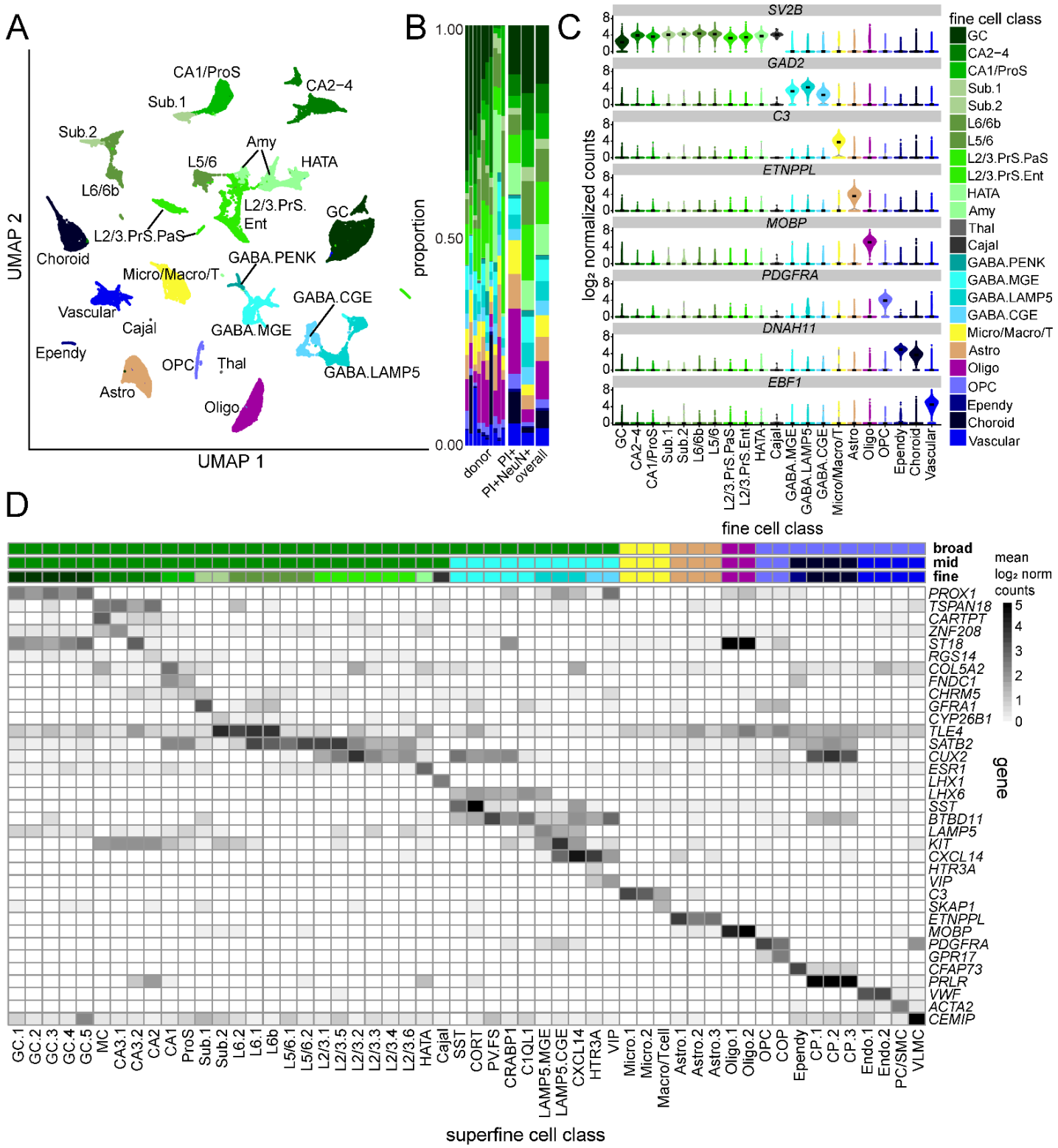

A
UMAP 2
UMAP 1
CA1/ProS
CA2−4
Sub.1
Sub.2
Amy
L5/6
HATA
L6/6b
L2/3.PrS.
Ent
GC
L2/3.PrS.PaS
Choroid
GABA.PENK
Micro/Macro/T
GABA.CGE
Vascular
Cajal
GABA.MGE
Ependy
OPC
Thal
GABA.LAMP5
Astro
Oligo
B
1.00
0.50
0.00
proportion
donor
PI+
PI+NeuN+
overall
C
SV2B
GAD2
C3
ETNPPL
MOBP
PDGFRA
DNAH11
EBF1
log2 normalized counts
fine cell class
GC
CA2−4
CA1/ProS
Sub.1
Sub.2
L6/6b
L5/6
L2/3.PrS.PaS
L2/3.PrS.Ent
HATA
Amy
Thal
Cajal
GABA.PENK
GABA.MGE
GABA.LAMP5
GABA.CGE
Micro/Macro/T
Astro
Oligo
OPC
Ependy
Choroid
Vascular
D
broad
mid
fine
mean log2 norm counts
gene
PROX1
TSPAN18
CARTPT
ZNF208
ST18
RGS14
COL5A2
FNDC1
CHRM5
GFRA1
CYP26B1
TLE4
SATB2
CUX2
ESR1
LHX1
LHX6
SST
BTBD11
LAMP5
KIT
CXCL14
HTR3A
VIP
C3
SKAP1
ETNPPL
MOBP
PDGFRA
GPR17
CFAP73
PRLR
VWF
ACTA2
CEMIP
GC.1
GC.2
GC.3
GC.4
GC.5
MC
CA3.1
CA3.2
CA2
CA1
ProS
Sub.1
Sub.2
L6.2
L6.1
L6b
L5/6.1
L5/6.2
L2/3.1
L2/3.5
L2/3.2
L2/3.3
L2/3.4
L2/3.6
HATA
Cajal
SST
CORT
PV.FS
CRABP1
C1QL1
LAMP5.MGE
LAMP5.CGE
CXCL14
HTR3A
VIP
Micro.1
Micro.2
Macro/Tcell
Astro.1
Astro.2
Astro.3
Oligo.1
Oligo.2
OPC
COP
Ependy
CP.1
CP.2
CP.3
Endo.1
Endo.2
PC/SMC
VLMC
superfine cell class

Fig. 4. Comparison of spot level cell type deconvolution algorithms.

(**A**) Spotplots of (i) counts for granule cell layer (GCL) marker gene *PPFIA2* which are compared to (ii) the proportion of granule cell (GC) marker genes with non zero expression as well as (iii) the counts of GCs as predicted by the three spot-level deconvolution algorithms for sample V12D07-332_B1 (Br3942).

(**B**) (i) raw immunofluorescence Visium-SPG image, followed by (ii) `Cellpose` masks colored by Classification And Regression Tree (`CART`) predicted broad cell class, (iii) followed by scatterpie spotplot of `RCTD` sub level predictions collapsed down to broad cell class.

(**C**) Comparison of counts for each cell type for each capture area between mid- and fine-level resolutions and collapsed onto the cell type resolution used by the `CART`, where values theoretically should precisely match.

(**D**) (i) Boxplots using GC cell type as an example, showing predicted counts for each method are averaged across all spots within each projected spatial domain for each Visium-SPG tissue section (ii) Distribution of fine level cell types across `PRECAST` spatial domains from all 3 tools on the Visium-H&E data. Note: *X*-axis is shared between (i) and (ii).

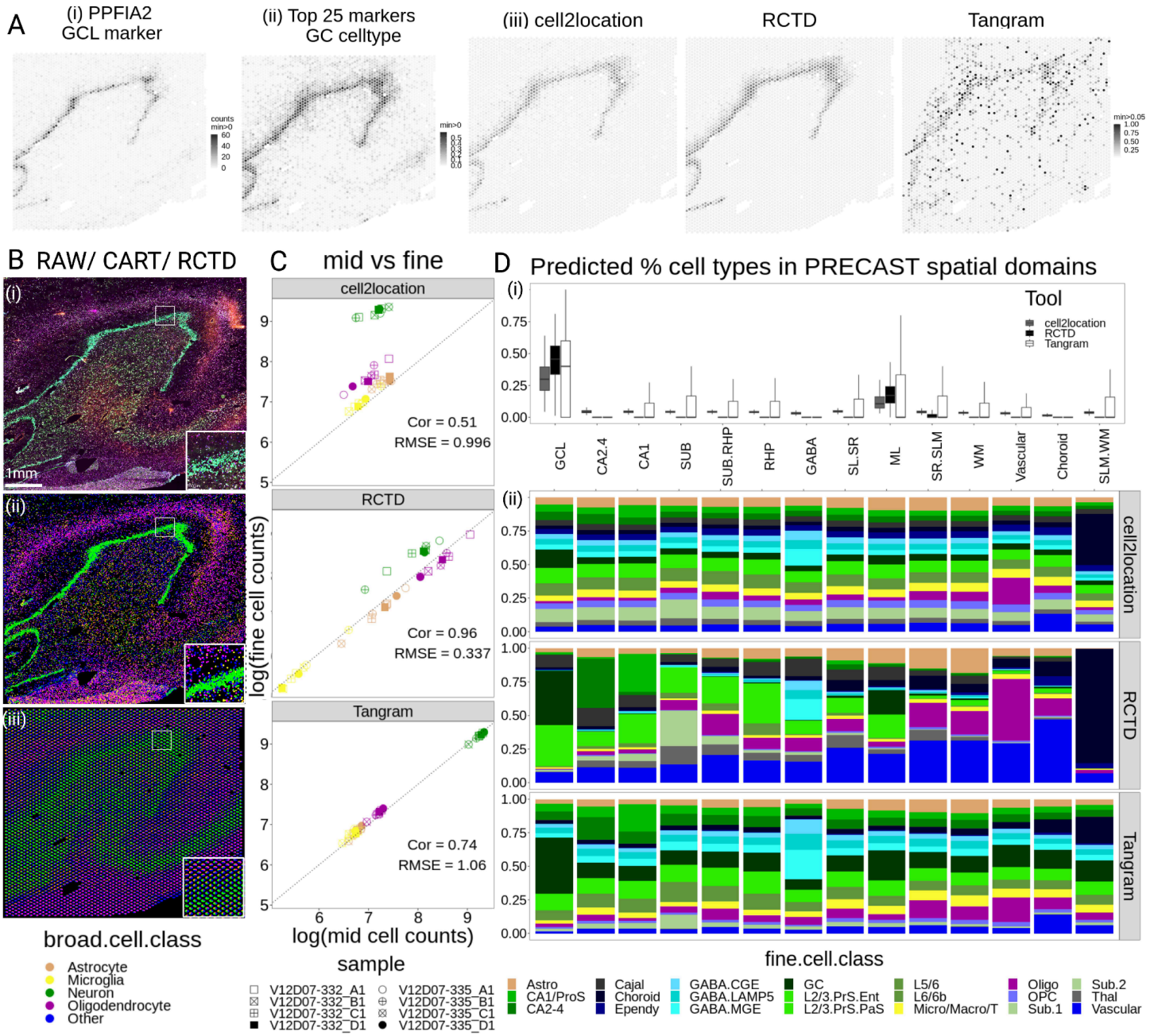

A
(i) PPFIA2 GCL marker
(ii) Top 25 markers GC celltype
(iii) cell2location
RCTD
Tangram
B RAW/ CART/ RCTD
(i)
(ii)
(iii)
1mm
C mid vs fine
cell2location
Cor = 0.51
RMSE = 0.996
RCTD
Cor = 0.96
RMSE = 0.337
Tangram
Cor = 0.74
RMSE = 1.06
log(fine cell counts)
log(mid cell counts)
D Predicted % cell types in PRECAST spatial domains
Tool
cell2location
RCTD
Tangram
GCL
CA2.4
CA1
SUB
SUB.RHP
RHP
GABA
SL.SR
ML
SR.SLM
WM
Vascular
Choroid
SLM.WM
broad.cell.class
Astrocyte
Microglia
Neuron
Oligodendrocyte
Other
sample
V12D07-332_A1
V12D07-332_B1
V12D07-332_C1
V12D07-332_D1
V12D07-335_A1
V12D07-335_B1
V12D07-335_C1
V12D07-335_D1
fine.cell.class
Astro
CA1/ProS
CA2-4
Cajal
Choroid
Ependy
GABA.CGE
GABA.LAMP5
GABA.MGE
GC
L2/3.PrS.Ent
L2/3.PrS.PaS
L5/6
L6/6b
Micro/Macro/T
Oligo
OPC
Sub.1
Sub.2
Thal
Vascular

Fig. 5. Non-negative matrix factorization reveals shared gene expression patterns between snRNA-seq and SRT data.

(**A**) Schematic of workflow for NMF-based analyses.

(**B**) UMAP representation of snRNA-seq data, binned into hexagonal bins of 100 nuclei each. Plot is colored by nuclei-level weights for nmf77.

(**C**) Boxplot showing nuclei-level weights for nmf77 across fine cell class groupings.

(**D**) Spot plot illustrating spatial distribution of predicted spot-level weights for nmf77 following transfer learning. Spot borders are colored by broad cell type, and spots are filled by predicted spot-level nmf77 weights.

(**E**) Hexbinned UMAP representation of snRNA-seq data colored by nuclei-level weights for nmf79.

(**F**) Boxplot showing nuclei-level weights for nmf79 across fine cell class groupings.

(**G**) Spot plot illustrating spatial distribution of predicted spot-level weights for nmf79 following transfer learning. Spot borders are colored by broad cell type, and spots are filled by predicted spot-level nmf79 weights.

(**H**) Hexbinned UMAP representation of snRNA-seq data colored by nuclei-level weights for nmf13.

(I) Boxplot showing nuclei-level weights for nmf13 across fine cell class groupings.

(**J**) Spot plot illustrating spatial distribution of predicted spot-level weights for nmf13 following transfer learning. Spot borders are colored by broad cell type, and spots are filled by predicted spot-level nmf13 weights.

(**K**) Heatmap of gene-level NMF weights of top marker genes for nmf77, nmf79, and nmf13.

(**L**) Plot showing enrichment of top GO terms for nmf77, nmf79, and nmf13 using results from over representation analysis of NMF patterns. Circles are colored by FDR-adjusted p-value and sized by the gene ratio for each term (**Methods**).

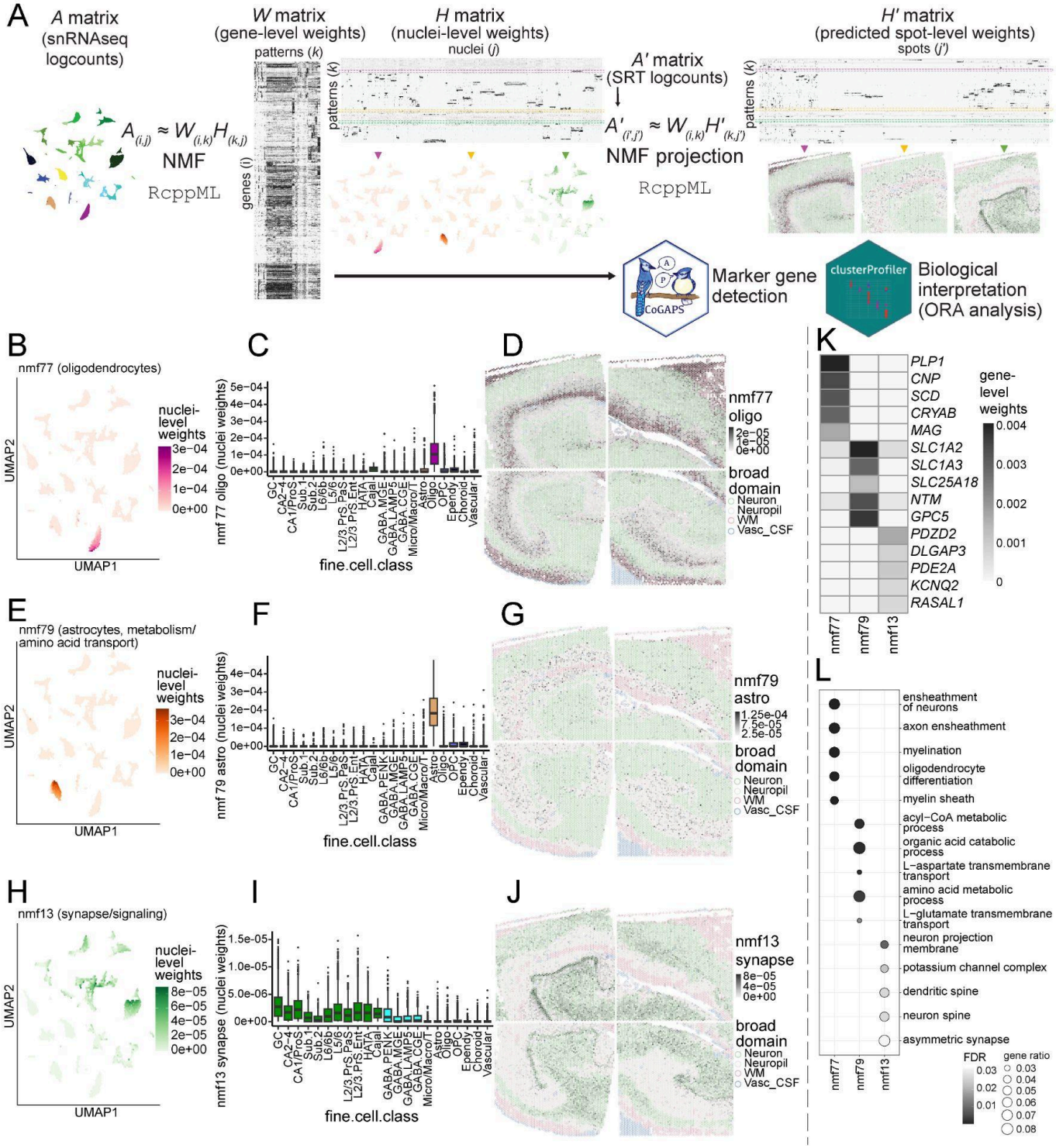

A
A matrix (snRNAseq logcounts)
W matrix (gene-level weights)
H matrix (nuclei-level weights)
patterns (k)
nuclei (j)
genes (i)
A′ matrix (SRT logcounts)
H′ matrix (predicted spot-level weights)
spots (j′)
A(i,j) ≈ W(i,k)H(k,j)
NMF
RcppML
A′(i′,j′) ≈ W(i,k)H′(k,j′)
NMF projection
RcppML
CoGAPS
Marker gene detection
clusterProfiler
Biological interpretation (ORA analysis)
B
nmf77 (oligodendrocytes)
nuclei-level weights
UMAP2
UMAP1
C
nmf 77 oligo (nuclei weights)
fine.cell.class
D
nmf77 oligo
broad domain
Neuron
Neuropil
WM
Vasc_CSF
E
nmf79 (astrocytes, metabolism/ amino acid transport)
F
nmf 79 astro (nuclei weights)
G
nmf79 astro
H
nmf13 (synapse/signaling)
I
nmf13 synapse (nuclei weights)
J
nmf13 synapse
K
gene-level weights
PLP1
CNP
SCD
CRYAB
MAG
SLC1A2
SLC1A3
SLC25A18
NTM
GPC5
PDZD2
DLGAP3
PDE2A
KCNQ2
RASAL1
nmf77
nmf79
nmf13
L
ensheathment of neurons
axon ensheathment
myelination
oligodendrocyte differentiation
myelin sheath
acyl−CoA metabolic process
organic acid catabolic process
L−aspartate transmembrane transport
amino acid metabolic process
L−glutamate transmembrane transport
neuron projection membrane
potassium channel complex
dendritic spine
neuron spine
asymmetric synapse
FDR
gene ratio

Fig. 6. NMF reveals a continuum of pyramidal cell types across HPC.

(**A**) Plot showing enrichment of pyramidal neuron-specific nuclei-level NMF weights (*x*-axis) across pyramidal neuron superfine cell classes (*y*-axis) in our human snRNA-seq data. Circles are colored by NMF weights for each pattern (aggregated by superfine cell class). Circles are sized by proportion of nuclei with nonzero weights for each pattern.

(**B**) Plot showing enrichment of predicted pyramidal neuron-specific spot-level NMF weights (*x*-axis) across pyramidal neuron cell body-rich spatial domains (*y*-axis) in our human SRT data. Circles are colored by predicted NMF weights for each pattern (aggregated by domain). Circles are sized by proportion of spots with nonzero predicted weights for each pattern.

(**C**) Heatmap of gene weights of top marker genes for each pyramidal neuron-specific NMF pattern.

(**D**) Spot plot illustrating spatial organization of patterns specific to CA1-4 neurons. Spot borders are colored to delineate between neuron cell body-rich spatial domains and other spatial domains. Spots are colored to reflect relative weights for each pattern within each spot.

(**E**) Sample as D, but illustrating patterns specific to CA1/ProS, Sub.1, Sub.2, and RHP L2/3.1 neurons.

(**F**) Plot showing enrichment of predicted pyramidal neuron-specific nuclei-level NMF weights (*x*-axis) across Allen Brain Atlas taxonomic cell types (*y*-axis) in mouse snmC-seq data [43]. Circles are colored by predicted NMF weights for each pattern (aggregated by domain). Circles are sized by proportion of nuclei with nonzero predicted weights for each pattern. See **Supplementary Table 2** for abbreviations.

(**G**) Same as F, but illustrating distribution of predicted NMF weights (*x*-axis) across neuron projection target region (*y*-axis).

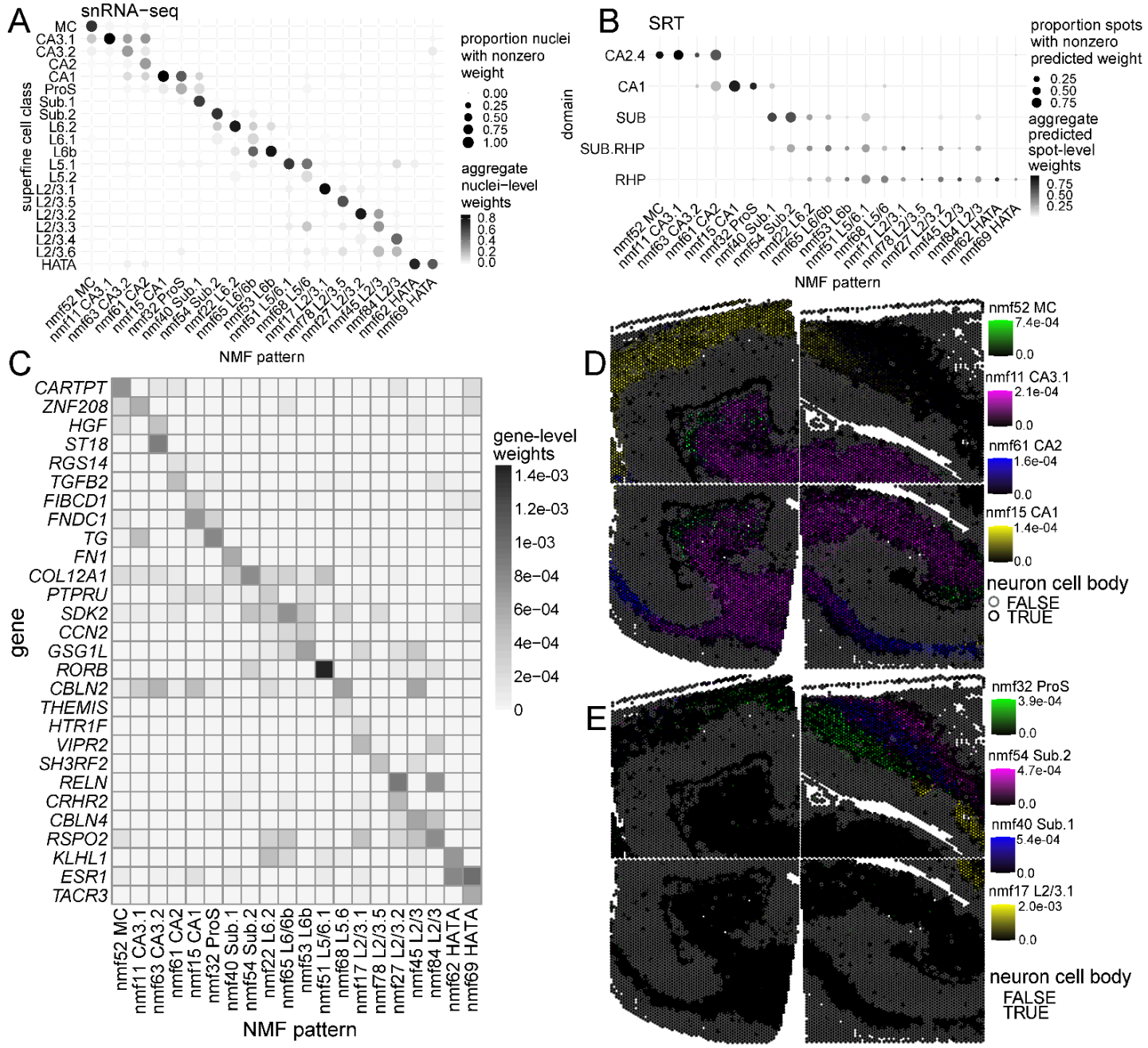

A
snRNA-seq
superfine cell class
MC
CA3.1
CA3.2
CA2
CA1
ProS
Sub.1
Sub.2
L6.2
L6.1
L6b
L5.1
L5.2
L2/3.1
L2/3.5
L2/3.2
L2/3.3
L2/3.4
L2/3.6
HATA
nmf52 MC
nmf11 CA3.1
nmf63 CA3.2
nmf61 CA2
nmf15 CA1
nmf32 ProS
nmf40 Sub.1
nmf54 Sub.2
nmf22 L6.2
nmf65 L6/6b
nmf53 L6b
nmf51 L5/6.1
nmf68 L5/6
nmf17 L2/3.1
nmf78 L2/3.5
nmf27 L2/3.2
nmf45 L2/3
nmf84 L2/3
nmf62 HATA
nmf69 HATA
NMF pattern
proportion nuclei with nonzero weight
0.00
0.25
0.50
0.75
1.00
aggregate nuclei-level weights
0.8
0.6
0.4
0.2
0.0
B
SRT
domain
CA2.4
CA1
SUB
SUB.RHP
RHP
proportion spots with nonzero predicted weight
0.25
0.50
0.75
aggregate predicted spot-level weights
0.75
0.50
0.25
C
gene
CARTPT
ZNF208
HGF
ST18
RGS14
TGFB2
FIBCD1
FNDC1
TG
FN1
COL12A1
PTPRU
SDK2
CCN2
GSG1L
RORB
CBLN2
THEMIS
HTR1F
VIPR2
SH3RF2
RELN
CRHR2
CBLN4
RSPO2
KLHL1
ESR1
TACR3
gene-level weights
1.4e-03
1.2e-03
1e-03
8e-04
6e-04
4e-04
2e-04
0
D
nmf52 MC
7.4e-04
0.0
nmf11 CA3.1
2.1e-04
0.0
nmf61 CA2
1.6e-04
0.0
nmf15 CA1
1.4e-04
0.0
neuron cell body
FALSE
TRUE
E
nmf32 ProS
3.9e-04
0.0
nmf54 Sub.2
4.7e-04
0.0
nmf40 Sub.1
5.4e-04
0.0
nmf17 L2/3.1
2.0e-03
0.0
neuron cell body
FALSE
TRUE
Pyramidal neuron NMF pattern projections, mouse HPC/RHP epi-retro-seq data (Zhou et al 2023)
F
Allen Institute subclass
CA3 Glut
CA2-FC-IG Glut
CA1-ProS Glut
SUB-ProS Glut
NP SUB Glut
CT SUB Glut
L6 IT CTX Glut
IT EP-CLA Glut
L6b/CT ENT Glut
L5 ET CTX Glut
L5/6 IT TPE-ENT Glut
L2/3 IT PPP Glut
L2 IT PPP-APr Glut
L2/3 IT ENT Glut
L2 IT ENT-po Glut
L2/3 IT PIR-ENTl Glut
ENTmv-PA-COAp Glut
COAp Grxcr2 Glut
aggregate predicted nuclei-level weights
0.75
0.50
0.25
0.00
proportion nuclei with nonzero weight
0.00
0.25
0.50
0.75
G
Target region
MOB
HY
ENT
AMY
STR
RSP
PTLp
HPF
TH
MOp
ACA
PIR
PFC
aggregate predicted nuclei-level weights
0.8
0.6
0.4
0.2
0.0
proportion nuclei with nonzero weight
0.0
0.1
0.2
0.3
0.4
0.5

Fig. 7: NMF identifies cell populations primed for recruitment in the human hippocampus.

(**A-E**) Plots showing relative weights of specific NMF patterns in ECS-activated HPC neurons (red) or control neurons (blue).

(**F-G**) Boxplots showing enrichment of granule cell-specific NMF patterns (nmf10 and nmf14, respectively) in granule cell superfine cell classes.

(**H**) Heatmap of gene-level NMF weights of top marker genes for nmf10, nmf14, nmf91, nmf20, nmf13.

(**I**) Spot plot illustrating spatial distribution of nmf10 and nmf14 following transfer learning. Spot borders are colored to delineate between neuron cell body-rich spatial domains and other spatial domains. Spots are colored to reflect relative weights for nmf10 (blue) and nmf14 (green) within each spot.

(**J**) Spot plot illustrating spatial distribution of nmf20, nmf13, and nmf91 following transfer learning. Spot borders are colored to delineate between neuron cell body-rich spatial domains and other spatial domains. Spots are colored to reflect relative weights for nmf20 (magenta), nmf13 (cyan), and nmf91 (yellow) within each spot.

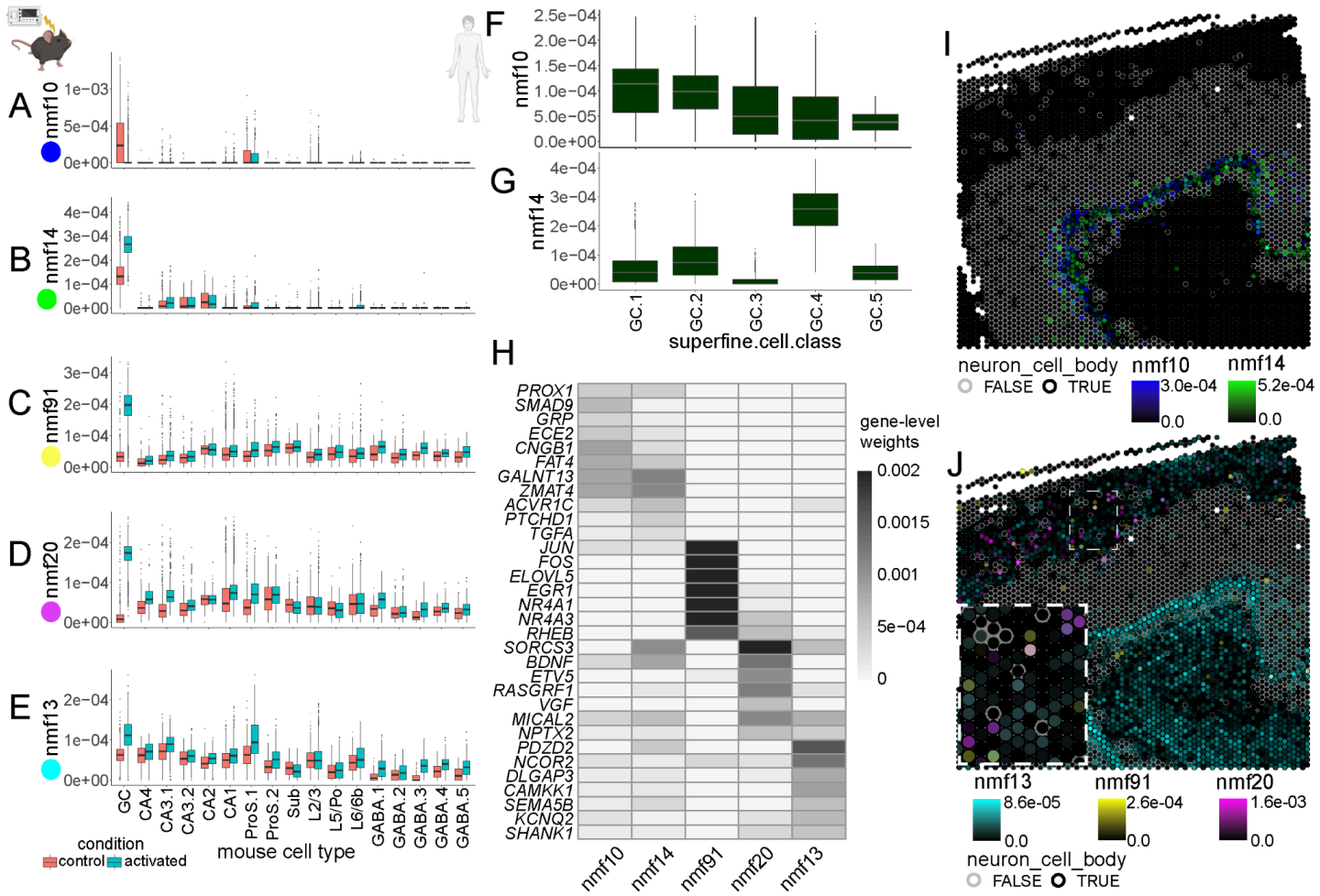
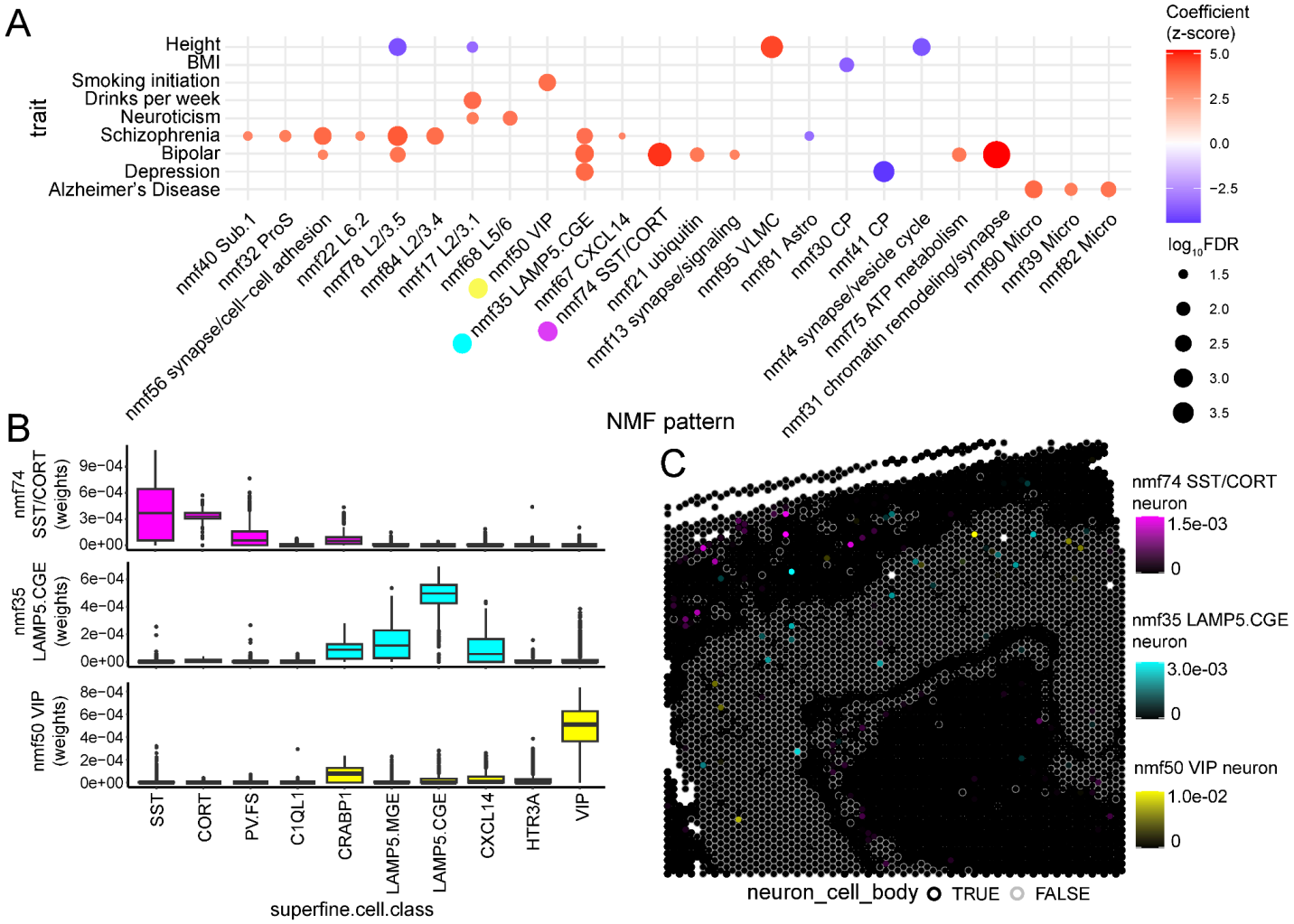


Fig. 8: Linkage disequilibrium score regression identifies differential disease risk heritability across NMF patterns.

(**A**) Plot showing LDSC regression coefficient z-scores for heritability of different polygenic traits (*y*-axis) across NMF patterns (*x*-axis). Coefficients found to be significantly different from the baseline model [50] following multiple testing correction (FDR < 0.05) are shown. Circles are colored by LDSC regression coefficient *z*-score and sized by -log10FDR. Positive coefficient *z*-scores denote a positive contribution of a given NMF pattern to trait heritability. Negative coefficient z-scores denote depletion of trait heritability within a given NMF pattern. Magenta, yellow, and cyan circles along *x*-axis indicate patterns highlighted in subsequent panels.

(**B**) Boxplots showing distribution of NMF pattern weights across snRNA-seq GABAergic cell classes for 3 patterns with positive contributions to disease heritability: nmf74 SST/CORT (magenta), nmf35 LAMP5.CGE (cyan), and nmf50 VIP (yellow).

(**C**) Spot plot illustrating spatial organization of the same 3 GABAergic neuron-specific patterns. Spot borders are colored to delineate between neuron cell body-rich spatial domains and neuron cell body-depleted spatial domains. Spots are colored to reflect relative weights for each pattern within each spot.



Supplementary Tables

Supplementary Table 1. Donor demographic information. Demographic information on brain donors, including Sample ID (BrNum), age at time of death, sex, psychiatric diagnosis, post-mortem interval (PMI), body mass index (BMI, calculated postmortem), Screening RNA Integrity Number (RIN, calculated at time of brain collection in the prefrontal cortex–PFC), and, for some samples, RIN measured in the tissue blocks included in this study.

Supplementary Table 2. **Abbreviations**. Guide to abbreviations used in naming spatial domains, snRNA-seq cell classes, and Allen Institute/BICCN nomenclature.

Supplementary Table 3. nnSVG gene ranks. Results from nnSVG analysis, used to identify spatially variable genes (SVGs) in the SRT data. Columns include ENSEMBL ID, gene name, overall gene rank, mean gene rank across all capture areas, number of capture areas for which each gene was highly ranked (within top 1000 most variable genes), and whether each gene was used for `PRECAST` clustering.

Supplementary Table 4. Differential expression results across pseudobulked spatial domains using the layer-enriched model. For each gene, a linear mixed-effects model was fit with counts aggregated (or pseudobulked) across spots within a spatial domain to identify differences in expression enriched in one domain compared to all other domains using Student's *t*-test statistics.

Supplementary Table 5. Differential expression results across four pseudobulked broad spatial domains using the layer-enriched model. Four broad domains: neuron-enriched domains (comprised of pyramidal, granule and GABAergic neuron clusters), neuropil-enriched domains (including ML, SL, SR, SLM.SGZ), WM, and vascular/cerebrospinal fluid (CSF)-related domains (comprised of the two domains mapping to vasculature and choroid plexus, respectively). For each gene, a linear mixed-effects model was fit with counts aggregated (or pseudobulked) across spots within a spatial domain to identify differences in expression enriched in one domain compared to all other domains using Student's *t*-test statistics.

Supplementary Table 6. Differential expression results across mid cell class labels using pairwise *t*-tests. Pairwise *t*-tests were performed between all pairs of mid cell classes for each gene. Columns include *p*-values, FDR adjusted *p*-values, summary $\log_2$ fold change, and $\log_2$ fold change for each pairwise comparison.

Supplementary Table 7. Differential expression results across fine cell class labels using pairwise *t*-tests. Pairwise *t*-tests were performed between all pairs of fine cell classes for each gene. Columns include *p*-values, FDR adjusted *p*-values, summary $\log_2$ fold change, and $\log_2$ fold change for each pairwise comparison.

Supplementary Table 8. Differential expression results across superfine cell class labels using pairwise *t*-tests. Pairwise *t*-tests were performed between all pairs of superfine cell classes for each gene. Columns include *p*-values, FDR adjusted *p*-values, summary $\log_2$ fold change, and $\log_2$ fold change for each pairwise comparison.

Supplementary Table 9. Differential expression results across pseudobulked "mid" cell class labels using the layer-enriched model. For each gene, a linear mixed-effects model was fit with counts aggregated (or pseudobulked) across nuclei within each mid-level cell class to identify differences in expression enriched in one mid-level cell class compared to all other mid-level cell classes using Student's *t*-test statistics.

Supplementary Table 10. Differential expression results across pseudobulked "fine" cell class labels using the layer-enriched model. For each gene, a linear mixed-effects model was fit with counts aggregated

(or pseudobulked) across nuclei within each fine-level cell class to identify differences in expression enriched in one fine-level cell class compared to all other fine-level cell classes using Student's *t*-test statistics.

Supplementary Table 11. Differential expression results across pseudobulked "superfine" cell class labels using the layer-enriched model. For each gene, a linear mixed-effects model was fit with counts aggregated (or pseudobulked) across nuclei within each superfine-level cell class to identify differences in expression enriched in one superfine-level cell class compared to all other superfine-level cell classes using Student's *t*-test statistics.

Supplementary Table 12. Spot deconvolution cell type marker gene statistics. For each cell type at mid and fine resolution, statistics of top 25 marker genes (mean ratio>1) extracted with the *DeconvoBuddies* package using the filtered (genes that overlapped with Visium data only) snRNAseq data.

Supplementary Table 13. RMSE and correlation stats. For each cell type at mid and fine resolution, Pearson Correlation and RMSE for spot-deconvolution software results compared with CART predictions were calculated for each Visum-SPG capture area.

Supplementary Table 14. **Gene-level NMF weights.** Table with gene-level weights for NMF patterns identified using `RcppML`. Each row is a gene and each column is an NMF pattern. Patterns were normalized such that each column sums to 1.

Supplementary Table 15. Marker genes for NMF patterns using `patternMarkers`. Results from NMF marker gene detection using `patternMarkers`. Each sheet contains a list of markers for a different pattern.

Supplementary Table 16. Gene Ontology (GO) over-representation analysis results for NMF patterns. Results from GO over-representation analysis. Each sheet contains results for a different pattern. Columns include GO ID, GO term description, p-values, FDR-corrected p-value, and gene ratio.

Bibliography

1. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;19: 534–546. doi:10.1038/s41592-022-01409-2
2. Moffitt JR, Lundberg E, Heyn H. The emerging landscape of spatial profiling technologies. Nat Rev Genet. 2022;23: 741–759. doi:10.1038/s41576-022-00515-3
3. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome Med. 2022;14: 68. doi:10.1186/s13073-022-01075-1
4. Maynard KR, Collado-Torres L, Weber LM, Uytingco C, Barry BK, Williams SR, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nat Neurosci. 2021;24: 425–436. doi:10.1038/s41593-020-00787-0
5. Huuki-Myers L, Spangler A, Eagles N, Montgomery KD, Kwon SH, Guo B, et al. Integrated single cell and unsupervised spatial transcriptomic analysis defines molecular anatomy of the human dorsolateral prefrontal cortex. BioRxiv. 2023. doi:10.1101/2023.02.15.528722
6. Weber LM, Divecha HR, Tran MN, Kwon SH, Spangler A, Montgomery KD, et al. The gene expression landscape of the human locus coeruleus revealed by single-nucleus and spatially-resolved transcriptomics. eLife. 2024;12. doi:10.7554/eLife.84628
7. Braun E, Danan-Gotthold M, Borm LE, Lee KW, Vinsland E, Lönnerberg P, et al. Comprehensive cell atlas of the first-trimester developing human brain. Science. 2023;382: eadf1226. doi:10.1126/science.adf1226
8. Strange BA, Witter MP, Lein ES, Moser EI. Functional organization of the hippocampal longitudinal axis. Nat Rev Neurosci. 2014;15: 655–669. doi:10.1038/nrn3785
9. Lisman J, Buzsáki G, Eichenbaum H, Nadel L, Ranganath C, Redish AD. Viewpoints: how the hippocampus contributes to memory, navigation and cognition. Nat Neurosci. 2017;20: 1434–1447. doi:10.1038/nn.4661
10. Ding S-L, Yao Z, Hirokawa KE, Nguyen TN, Graybuck LT, Fong O, et al. Distinct Transcriptomic Cell Types and Neural Circuits of the Subiculum and Prosubiculum along the Dorsal-Ventral Axis. Cell Rep. 2020;31: 107648. doi:10.1016/j.celrep.2020.107648
11. Cembrowski MS, Wang L, Lemire AL, Copeland M, DiLisio SF, Clements J, et al. The subiculum is a patchwork of discrete subregions. eLife. 2018;7. doi:10.7554/eLife.37701
12. Yao Z, van Velthoven CTJ, Nguyen TN, Goldy J, Sedeno-Cortes AE, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell. 2021;184: 3222-3241.e26. doi:10.1016/j.cell.2021.04.021
13. Qiu S, Hu Y, Huang Y, Gao T, Wang X, Wang D, et al. Whole-brain spatial organization of hippocampal single-neuron projectomes. Science. 2024;383: eadj9198. doi:10.1126/science.adj9198
14. Ayhan F, Kulkarni A, Berto S, Sivaprakasam K, Douglas C, Lega BC, et al. Resolving cellular and molecular diversity along the hippocampal anterior-to-posterior axis in humans. Neuron. 2021;109: 2091-2105.e6. doi:10.1016/j.neuron.2021.05.003
15. Franjic D, Skarica M, Ma S, Arellano JI, Tebbenkamp ATN, Choi J, et al. Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. Neuron. 2022;110: 452-469.e14. doi:10.1016/j.neuron.2021.10.036
16. Glock C, Biever A, Tushev G, Nassim-Assir B, Kao A, Bartnik I, et al. The translatome of neuronal cell

bodies, dendrites, and axons. Proc Natl Acad Sci USA. 2021;118. doi:10.1073/pnas.2113929118

17. Liu W, Liao X, Luo Z, Yang Y, Lau MC, Jiao Y, et al. Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. Nat Commun. 2023;14: 296. doi:10.1038/s41467-023-35947-w

18. Weber LM, Saha A, Datta A, Hansen KD, Hicks SC. nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. Nat Commun. 2023;14: 4059. doi:10.1038/s41467-023-39748-z

19. Lee M, Yoon J, Song H, Lee B, Lam DT, Yoon J, et al. Tcf7l2 plays crucial roles in forebrain development through regulation of thalamic and habenular neuron identity and connectivity. Dev Biol. 2017;424: 62–76. doi:10.1016/j.ydbio.2017.02.010

20. Nagalski A, Puelles L, Dabrowski M, Wegierski T, Kuznicki J, Wisniewska MB. Molecular anatomy of the thalamic complex and the underlying transcription factors. Brain Struct Funct. 2016;221: 2493–2510. doi:10.1007/s00429-015-1052-5

21. Tran MN, Maynard KR, Spangler A, Huuki LA, Montgomery KD, Sadashivaiah V, et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. Neuron. 2021;109: 3088-3103.e5. doi:10.1016/j.neuron.2021.09.001

22. Letts VA, Mahaffey CL, Beyer B, Frankel WN. A targeted mutation in Cacng4 exacerbates spike-wave seizures in stargazer (Cacng2) mice. Proc Natl Acad Sci USA. 2005;102: 2123–2128. doi:10.1073/pnas.0409527102

23. Mansour A, Khachaturian H, Lewis ME, Akil H, Watson SJ. Autoradiographic differentiation of mu, delta, and kappa opioid receptors in the rat forebrain and midbrain. J Neurosci. 1987;7: 2445–2464.

24. Cembrowski MS, Wang L, Sugino K, Shields BC, Spruston N. Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons. eLife. 2016;5: e14997. doi:10.7554/eLife.14997

25. Long Y, Ang KS, Li M, Chong KLK, Sethi R, Zhong C, et al. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. Nat Commun. 2023;14: 1155. doi:10.1038/s41467-023-36796-3

26. Li B, Zhang W, Guo C, Xu H, Li L, Fang M, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. Nat Methods. 2022;19: 662–670. doi:10.1038/s41592-022-01480-9

27. Kleshchevnikov V, Shmatko A, Dann E, Aivazidis A, King HW, Li T, et al. Cell2location maps fine-grained cell types in spatial transcriptomics. Nat Biotechnol. 2022;40: 661–671. doi:10.1038/s41587-021-01139-4

28. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40: 517–526. doi:10.1038/s41587-021-00830-w

29. Biancalani T, Scalia G, Buffoni L, Avasthi R, Lu Z, Sanger A, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nat Methods. 2021;18: 1352–1362. doi:10.1038/s41592-021-01264-7

30. Huuki-Myers LA, Montgomery KD, Kwon SH, Cinquemani S, Maden SK, Eagles NJ, et al. Benchmark of cellular deconvolution methods using a multi-assay reference dataset from postmortem human prefrontal cortex. BioRxiv. 2024. doi:10.1101/2024.02.09.579665

31. Stein-O'Brien GL, Arora R, Culhane AC, Favorov AV, Garmire LX, Greene CS, et al. Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet. 2018;34: 790–805. doi:10.1016/j.tig.2018.07.003

32. DeBruine ZJ, Melcher K, Triche TJ. Fast and robust non-negative matrix factorization for single-cell experiments. BioRxiv. 2021. doi:10.1101/2021.09.01.458620

33. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. eLife. 2019;8. doi:10.7554/eLife.43803

34. Stein-O'Brien GL, Clark BS, Sherman T, Zibetti C, Hu Q, Sealfon R, et al. Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. Cell Syst. 2019;8: 395-411.e8. doi:10.1016/j.cels.2019.04.004

35. Olney KC, Todd KT, Pallegar PN, Jensen TD, Cadiz MP, Gibson KA, et al. Widespread choroid plexus contamination in sampling and profiling of brain tissue. Mol Psychiatry. 2022;27: 1839–1847. doi:10.1038/s41380-021-01416-3

36. Stein-O'Brien GL, Carey JL, Lee WS, Considine M, Favorov AV, Flam E, et al. PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics. 2017;33: 1892–1894. doi:10.1093/bioinformatics/btx058

37. The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49: D325–D334. doi:10.1093/nar/gkaa1113

38. Földy C, Darmanis S, Aoto J, Malenka RC, Quake SR, Südhof TC. Single-cell RNAseq reveals cell adhesion molecule profiles in electrophysiologically defined neurons. Proc Natl Acad Sci USA. 2016;113: E5222-31. doi:10.1073/pnas.1610155113

39. Langlieb J, Sachdev NS, Balderrama KS, Nadaf NM, Raj M, Murray E, et al. The molecular cytoarchitecture of the adult mouse brain. Nature. 2023;624: 333–342. doi:10.1038/s41586-023-06818-7

40. Zhang M, Pan X, Jung W, Halpern AR, Eichhorn SW, Lei Z, et al. Molecularly defined and spatially resolved cell atlas of the whole mouse brain. Nature. 2023;624: 343–354. doi:10.1038/s41586-023-06808-9

41. Yao Z, van Velthoven CTJ, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624: 317–332. doi:10.1038/s41586-023-06812-z

42. Wheeler DW, Kopsick JD, Sutton N, Tecuatl C, Komendantov AO, Nadella K, et al. Hippocampome.org 2.0 is a knowledge base enabling data-driven spiking neural network simulations of rodent hippocampal circuits. eLife. 2024;12. doi:10.7554/eLife.90597

43. Zhou J, Zhang Z, Wu M, Liu H, Pang Y, Bartlett A, et al. Brain-wide correspondence of neuronal epigenomics and distant projections. Nature. 2023;624: 355–365. doi:10.1038/s41586-023-06823-w

44. Nelson ED, Maynard KR, Nicholas KR, Tran MN, Divecha HR, Collado-Torres L, et al. Activity-regulated gene expression across cell types of the mouse hippocampus. Hippocampus. 2023;33: 1009–1027. doi:10.1002/hipo.23548

45. Breiderhoff T, Christiansen GB, Pallesen LT, Vaegter C, Nykjaer A, Holm MM, et al. Sortilin-related receptor SORCS3 is a postsynaptic modulator of synaptic depression and fear extinction. PLoS ONE. 2013;8: e75006. doi:10.1371/journal.pone.0075006

46. Shi H-J, Wang S, Wang X-P, Zhang R-X, Zhu L-J. Hippocampus: molecular, cellular, and circuit features in anxiety. Neurosci Bull. 2023;39: 1009–1026. doi:10.1007/s12264-023-01020-1

47. Small SA, Schobel SA, Buxton RB, Witter MP, Barnes CA. A pathophysiological framework of hippocampal dysfunction in ageing and disease. Nat Rev Neurosci. 2011;12: 585–601. doi:10.1038/nrn3085

48. Kohen R, Dobra A, Tracy JH, Haugen E. Transcriptome profiling of human hippocampus dentate gyrus granule cells in mental illness. Transl Psychiatry. 2014;4: e366. doi:10.1038/tp.2014.9

49. Jaffe AE, Hoeppner DJ, Saito T, Blanpain L, Ukaigwe J, Burke EE, et al. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. Nat Neurosci. 2020;23: 510–519. doi:10.1038/s41593-020-0604-z

50. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228–1235. doi:10.1038/ng.3404

51. Bulik-Sullivan B, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47: 291–295. doi:10.1038/ng.3211

52. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018;50: 621–629. doi:10.1038/s41588-018-0081-4

53. Demontis D, Walters GB, Athanasiadis G, Walters R, Therrien K, Nielsen TT, et al. Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. Nat Genet. 2023;55: 198–208. doi:10.1038/s41588-022-01285-8

54. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet. 2022;54: 412–436. doi:10.1038/s41588-022-01024-z

55. Watson HJ, Yilmaz Z, Thornton LM, Hübel C, Coleman JRI, Gaspar HA, et al. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. Nat Genet. 2019;51: 1207–1214. doi:10.1038/s41588-019-0439-2

56. Huang J, Huffman JE, Huang Y, Do Valle Í, Assimes TL, Raghavan S, et al. Genomics and phenomics of body mass index reveals a complex disease network. Nat Commun. 2022;13: 7973. doi:10.1038/s41467-022-35553-2

57. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021;53: 817–829. doi:10.1038/s41588-021-00857-4

58. Saunders GRB, Wang X, Chen F, Jang S-K, Liu M, Wang C, et al. Genetic diversity fuels gene discovery for tobacco and alcohol use. Nature. 2022;612: 720–724. doi:10.1038/s41586-022-05477-4

59. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610: 704–712. doi:10.1038/s41586-022-05275-y

60. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51: 431–444. doi:10.1038/s41588-019-0344-8

61. Howard DM, Adams MJ, Clarke T-K, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22: 343–352. doi:10.1038/s41593-018-0326-7

62. International League Against Epilepsy Consortium on Complex Epilepsies. GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. Nat Genet. 2023;55: 1471–1482. doi:10.1038/s41588-023-01485-w

63. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022;604: 502–508. doi:10.1038/s41586-022-04434-5

64. Nagel M, Jansen PR, Stringer S, Watanabe K, de Leeuw CA, Bryois J, et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. Nat Genet. 2018;50: 920–927. doi:10.1038/s41588-018-0151-7

65. Nalls MA, Blauwendraat C, Vallerga CL, Heilbron K, Bandres-Ciga S, Chang D, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet Neurol. 2019;18: 1091–1102. doi:10.1016/S1474-4422(19)30320-5

66. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018;50: 906–908. doi:10.1038/s41588-018-0144-6

67. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, de Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet. 2018;50: 912–919. doi:10.1038/s41588-018-0152-6

68. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 2012;44: 981–990. doi:10.1038/ng.2383

69. Schmidt MJ, Mirnics K. Neurodevelopment, GABA system dysfunction, and schizophrenia. Neuropsychopharmacology. 2015;40: 190–206. doi:10.1038/npp.2014.95

70. Klausberger T, Somogyi P. Neuronal diversity and temporal dynamics: the unity of hippocampal circuit operations. Science. 2008;321: 53–57. doi:10.1126/science.1149381

71. Topolnik L, Tamboli S. The role of inhibitory circuits in hippocampal memory processing. Nat Rev Neurosci. 2022;23: 476–492. doi:10.1038/s41583-022-00599-0

72. Rue-Albrecht K, Marini F, Soneson C, Lun ATL. iSEE: Interactive SummarizedExperiment Explorer. [version 1; peer review: 3 approved]. F1000Res. 2018;7: 741. doi:10.12688/f1000research.14966.1

73. Sriworarat C, Nguyen A, Eagles NJ, Collado-Torres L, Martinowich K, Maynard KR, et al. Performant web-based interactive visualization tool for spatially-resolved transcriptomics experiments. Biol Imaging. 2023; 1–10. doi:10.1017/S2633903X2300017X

74. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41: D991-5. doi:10.1093/nar/gks1193

75. Wee RWS, MacAskill AF. Biased Connectivity of Brain-wide Inputs to Ventral Subiculum Output Neurons. Cell Rep. 2020;30: 3644-3654.e6. doi:10.1016/j.celrep.2020.02.093

76. Gergues MM, Han KJ, Choi HS, Brown B, Clausing KJ, Turner VS, et al. Circuit and molecular architecture of a ventral hippocampal network. Nat Neurosci. 2020;23: 1444–1452.

doi:10.1038/s41593-020-0705-8

77. Hallock HL, Quillian HM, Maynard KR, Mai Y, Chen H-Y, Hamersky GR, et al. Molecularly defined hippocampal inputs regulate population dynamics in the prelimbic cortex to suppress context fear memory retrieval. Biol Psychiatry. 2020;88: 554–565. doi:10.1016/j.biopsych.2020.04.014

78. Jimenez JC, Su K, Goldberg AR, Luna VM, Biane JS, Ordek G, et al. Anxiety Cells in a Hippocampal-Hypothalamic Circuit. Neuron. 2018;97: 670-683.e6. doi:10.1016/j.neuron.2018.01.016

79. Thompson SM, Kallarackal AJ, Kvarta MD, Van Dyke AM, LeGates TA, Cai X. An excitatory synapse hypothesis of depression. Trends Neurosci. 2015;38: 279–294. doi:10.1016/j.tins.2015.03.003

80. Lieberman JA, Girgis RR, Brucato G, Moore H, Provenzano F, Kegeles L, et al. Hippocampal dysfunction in the pathophysiology of schizophrenia: a selective review and hypothesis for early detection and intervention. Mol Psychiatry. 2018;23: 1764–1772. doi:10.1038/mp.2017.249

81. Rowland DC, Obenhaus HA, Skytøen ER, Zhang Q, Kentros CG, Moser EI, et al. Functional properties of stellate cells in medial entorhinal cortex layer II. eLife. 2018;7. doi:10.7554/eLife.36664

82. Schobel SA, Kelly MA, Corcoran CM, Van Heertum K, Seckinger R, Goetz R, et al. Anterior hippocampal and orbitofrontal cortical structural brain abnormalities in association with cognitive deficits in schizophrenia. Schizophr Res. 2009;114: 110–118. doi:10.1016/j.schres.2009.07.016

83. Price RB, Duman R. Neuroplasticity in cognitive and psychological mechanisms of depression: an integrative model. Mol Psychiatry. 2020;25: 530–543. doi:10.1038/s41380-019-0615-x

84. Duman RS, Sanacora G, Krystal JH. Altered connectivity in depression: GABA and glutamate neurotransmitter deficits and reversal by novel treatments. Neuron. 2019;102: 75–90. doi:10.1016/j.neuron.2019.03.013

85. Johnson KA, Okun MS, Scangos KW, Mayberg HS, de Hemptinne C. Deep brain stimulation for refractory major depressive disorder: a comprehensive review. Mol Psychiatry. 2024. doi:10.1038/s41380-023-02394-4

86. Mayberg HS, Lozano AM, Voon V, McNeely HE, Seminowicz D, Hamani C, et al. Deep brain stimulation for treatment-resistant depression. Neuron. 2005;45: 651–660. doi:10.1016/j.neuron.2005.02.014

87. Mayberg HS. Targeted electrode-based modulation of neural circuits for depression. J Clin Invest. 2009;119: 717–725. doi:10.1172/JCI38454

88. Ma H, Khaled HG, Wang X, Mandelberg NJ, Cohen SM, He X, et al. Excitation-transcription coupling, neuronal gene expression and synaptic plasticity. Nat Rev Neurosci. 2023;24: 672–692. doi:10.1038/s41583-023-00742-5

89. Pernía-Andrade AJ, Jonas P. Theta-gamma-modulated synaptic currents in hippocampal granule cells in vivo define a mechanism for network oscillations. Neuron. 2014;81: 140–152. doi:10.1016/j.neuron.2013.09.046

90. Senzai Y, Buzsáki G. Physiological properties and behavioral correlates of hippocampal granule cells and mossy cells. Neuron. 2017;93: 691-704.e5. doi:10.1016/j.neuron.2016.12.011

91. Diamantaki M, Frey M, Berens P, Preston-Ferrer P, Burgalossi A. Sparse activity of identified dentate granule cells during spatial exploration. eLife. 2016;5. doi:10.7554/eLife.20252

92. Borzello M, Ramirez S, Treves A, Lee I, Scharfman H, Stark C, et al. Assessments of dentate gyrus function: discoveries and debates. Nat Rev Neurosci. 2023;24: 502–517.

doi:10.1038/s41583-023-00710-z

93. Erwin SR, Sun W, Copeland M, Lindo S, Spruston N, Cembrowski MS. A Sparse, Spatially Biased Subtype of Mature Granule Cell Dominates Recruitment in Hippocampal-Associated Behaviors. Cell Rep. 2020;31: 107551. doi:10.1016/j.celrep.2020.107551

94. Ohline SM, Wake KL, Hawkridge MV, Dinnunhan MF, Hegemann RU, Wilson A, et al. Adult-born dentate granule cell excitability depends on the interaction of neuron age, ontogenetic age and experience. Brain Struct Funct. 2018;223: 3213–3228. doi:10.1007/s00429-018-1685-2

95. Marín-Burgin A, Mongiat LA, Pardi MB, Schinder AF. Unique processing during a period of high excitation/inhibition balance in adult-born neurons. Science. 2012;335: 1238–1242. doi:10.1126/science.1214956

96. McHugh SB, Lopes-Dos-Santos V, Gava GP, Hartwich K, Tam SKE, Bannerman DM, et al. Adult-born dentate granule cells promote hippocampal population sparsity. Nat Neurosci. 2022;25: 1481–1491. doi:10.1038/s41593-022-01176-5

97. Kirschen GW, Shen J, Tian M, Schroeder B, Wang J, Man G, et al. Active Dentate Granule Cells Encode Experience to Promote the Addition of Adult-Born Hippocampal Neurons. J Neurosci. 2017;37: 4661–4678. doi:10.1523/JNEUROSCI.3417-16.2017

98. Denny CA, Kheirbek MA, Alba EL, Tanaka KF, Brachman RA, Laughman KB, et al. Hippocampal memory traces are differentially modulated by experience, time, and adult neurogenesis. Neuron. 2014;83: 189–201. doi:10.1016/j.neuron.2014.05.018

99. Parylak SL, Qiu F, Linker SB, Gallina IS, Lim CK, Preciado D, et al. Neuronal activity-related transcription is blunted in immature compared to mature dentate granule cells. Hippocampus. 2023;33: 412–423. doi:10.1002/hipo.23515

100. Targa Dias Anastacio H, Matosin N, Ooi L. Neuronal hyperexcitability in Alzheimer's disease: what are the drivers behind this aberrant phenotype? Transl Psychiatry. 2022;12: 257. doi:10.1038/s41398-022-02024-7

101. Setti SE, Hunsberger HC, Reed MN. Alterations in hippocampal activity and alzheimer's disease. Trans Issues Psychol Sci. 2017;3: 348–356. doi:10.1037/tps0000124

102. Busche MA, Chen X, Henning HA, Reichwald J, Staufenbiel M, Sakmann B, et al. Critical role of soluble amyloid-β for early hippocampal hyperactivity in a mouse model of Alzheimer's disease. Proc Natl Acad Sci USA. 2012;109: 8740–8745. doi:10.1073/pnas.1206171109

103. Kazim SF, Chuang S-C, Zhao W, Wong RKS, Bianchi R, Iqbal K. Early-Onset Network Hyperexcitability in Presymptomatic Alzheimer's Disease Transgenic Mice Is Suppressed by Passive Immunization with Anti-Human APP/Aβ Antibody and by mGluR5 Blockade. Front Aging Neurosci. 2017;9: 71. doi:10.3389/fnagi.2017.00071

104. Lipska BK, Deep-Soboslay A, Weickert CS, Hyde TM, Martin CE, Herman MM, et al. Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. Biol Psychiatry. 2006;60: 650–658. doi:10.1016/j.biopsych.2006.06.019

105. 10x Genomics. spaceranger count, 10x Genomics. [cited 17 Apr 2022]. Available: https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/using/count

106. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020;17: 137–145. doi:10.1038/s41592-019-0654-x

107. Righelli D, Weber LM, Crowell HL, Pardo B, Collado-Torres L, Ghazanfar S, et al. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. Bioinformatics. 2022;38: 3128–3131. doi:10.1093/bioinformatics/btac299

108. Pardo B, Spangler A, Weber LM, Page SC, Hicks SC, Jaffe AE, et al. spatialLIBD: an R/Bioconductor package to visualize spatially-resolved transcriptomics data. BMC Genomics. 2022;23: 434. doi:10.1186/s12864-022-08601-w

109. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. [version 2; peer review: 3 approved, 2 approved with reservations]. F1000Res. 2016;5: 2122. doi:10.12688/f1000research.9501.2

110. McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33: 1179–1186. doi:10.1093/bioinformatics/btw777

111. Ly CV, Verstreken P. Mitochondria at the synapse. Neuroscientist. 2006;12: 291–299. doi:10.1177/1073858406287661

112. Devine MJ, Kittler JT. Mitochondria at the neuronal presynapse in health and disease. Nat Rev Neurosci. 2018;19: 63–80. doi:10.1038/nrn.2017.170

113. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 2018. doi:10.48550/arxiv.1802.03426

114. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15: 343–346. doi:10.1038/nmeth.4636

115. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184: 3573-3587. doi:10.1016/j.cell.2021.04.048

116. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16: 1289–1296. doi:10.1038/s41592-019-0619-0

117. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26: 139–140. doi:10.1093/bioinformatics/btp616

118. Lun ATL, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. Methods Mol Biol. 2016;1418: 391–416. doi:10.1007/978-1-4939-3578-9_19

119. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. doi:10.1093/nar/gkv007

120. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

121. Amaral DG, Scharfman HE, Lavenex P. The dentate gyrus: fundamental neuroanatomical organization (dentate gyrus for dummies). Prog Brain Res. 2007;163: 3–22. doi:10.1016/S0079-6123(07)63001-5

122. Tippani M, Divecha HR, Catallini JL, Kwon SH, Weber LM, Spangler A, et al. *VistoSeg* : processing utilities for high resolution images for spatially resolved transcriptomics data. Biol Imaging. 2023; 1–21.

doi:10.1017/S2633903X23000235

123. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2: e453. doi:10.7717/peerj.453

124. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas Jamboree, et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. 2019;20: 63. doi:10.1186/s13059-019-1662-y

125. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019;573: 61–68. doi:10.1038/s41586-019-1506-7

126. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019;20: 295. doi:10.1186/s13059-019-1861-6

127. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36: 421–427. doi:10.1038/nbt.4091

128. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17: 75. doi:10.1186/s13059-016-0947-7

129. Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. InterJournal Complex Systems. 2005;1695.

130. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008: P10008. doi:10.1088/1742-5468/2008/10/P10008

131. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12: 2825–2830.

132. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401: 788–791. doi:10.1038/44565

133. Owen AB, Perry PO. Bi-cross-validation of the SVD and the nonnegative matrix factorization. Ann Appl Stat. 2009;3: 564–594. doi:10.1214/08-AOAS227

134. Fertig EJ, Ding J, Favorov AV, Parmigiani G, Ochs MF. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics. 2010;26: 2792–2793. doi:10.1093/bioinformatics/btq503

135. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2: 100141. doi:10.1016/j.xinn.2021.100141

136. Myre JM, Frahm E, Lilja DJ, Saar MO. TNT-NN: A Fast Active Set Method for Solving Large Non-Negative Least Squares Problems. Procedia Computer Science. 2017;108: 755–764. doi:10.1016/j.procs.2017.05.194

137. Zhang B, He P, Lawrence JEG, Wang S, Tuck E, Williams BA, et al. A human embryonic limb cell atlas resolved in space and time. Nature. 2023. doi:10.1038/s41586-023-06806-x

138. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9: 676–682. doi:10.1038/nmeth.2019

139. TrakEM2. [cited 26 Feb 2024]. Available: https://imagej.net/plugins/trakem2/

140. Cardona A, Saalfeld S, Schindelin J, Arganda-Carreras I, Preibisch S, Longair M, et al. TrakEM2

software for neural circuit reconstruction. PLoS ONE. 2012;7: e38011.
doi:10.1371/journal.pone.0038011