Modular cell type organization of cortical areas revealed by in situ sequencing

Xiaoyin Chen[†*,1], Stephan Fischer[†,3,4], Aixin Zhang[1], Jesse Gillis*[,2], Anthony M Zador*[,3]

†These authors contributed equally to this work
*Correspondence: xiaoyin.chen@alleninstitute.org, jesse.gillis@utoronto.ca, zador@cshl.edu

1. Allen Institute for Brain Science, Seattle, WA, USA

2. University of Toronto, Toronto, Ontario, Canada

3. Cold Spring Harbor Laboratory, New York, NY, USA

4. Current address: Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France

Abstract

The cortex is composed of neuronal types with diverse gene expression that are organized into specialized cortical areas. These areas, each with characteristic cytoarchitecture [1-3], connectivity[4,5], and neuronal activity [6-10], are wired into modular networks[4,5,11]. However, it remains unclear whether cortical areas and their modular organization can be similarly defined by their transcriptomic signatures. Here we used BARseq, a high-throughput in situ sequencing technique, to interrogate the expression of 107 cell type marker genes in 1.2 million cells over a mouse forebrain hemisphere at cellular resolution. *De novo* clustering of gene expression in single neurons revealed transcriptomic types that were consistent with previous single-cell RNAseq studies[12,13]. Within medium-grained cell types that are shared across all cortical areas, gene expression and the distribution of fine-grained cell types vary along the contours of cortical areas. The compositions of transcriptomic types are highly predictive of cortical area identity. We grouped cortical areas into modules so that areas within a module, but not across modules, had similar compositions of transcriptomic types. Strikingly, these modules match cortical subnetworks that are highly interconnected [4,5,11], suggesting that cortical areas that are similar in cell types are also wired together. This "wire-by-similarity" rule reflects a novel organizing principle for the connectivity of cortical areas. Our BARseq-based strategy is high-throughput and low-cost, and scaling up this approach to many animals can potentially reveal the brain-wide molecular architecture across individuals, developmental times, and disease models.

Main text

The vertebrate brain is organized into subregions that are functionally specialized and anatomically distinct. This spatial specialization in function and structure is established by developmental processes involving intrinsic genetic programs and/or external signaling [14]. Both intrinsic and extrinsic developmental processes drive the expression of specific sets of genes, which together specify the correct cell fates and establish specialized cellular properties [15]. Although gene expression can change during cell maturation and remains dynamic in response to internal cellular conditions and external stimuli, a core transcriptional program that maintains the cellular identity usually remains steady in mature neurons [16]. Thus, resolving the expression of core sets of genes that distinguish different types of neurons provides insight into the functional and structural specialization of neurons.

Many large brain structures are spatially organized into divisions, or modules, within which neurons are more similar in morphology, connectivity, and activity. In the cortex, these modules usually involve a set of adjacent cortical areas that are highly interconnected [4,5,11] and correlated in neuronal activity [6-10]. This modular organization is perturbed in Alzheimer's disease[17], mild cognitive impairment[18], schizophrenia [19], and depression [20], suggesting that cortical modules are critical to normal functions. In contrast to this modular organization in activity and connectivity, many cortical areas share the same medium-grained and fine-grained transcriptomically defined neuronal types [13,21]. Whether and how the modular organization of cortical connectivity and activity is reflected in the transcriptomic signatures of areas is unknown.

To address this question, here we apply BARseq [22,23] to interrogate gene expression and the distribution of excitatory neuron types across a mouse forebrain hemisphere at high spatial resolution. BARseq is a form of *in situ* sequencing [24-26], in which Illumina sequencing-by-synthesis chemistry is used to achieve a robust readout of both endogenous mRNAs and synthetic RNA barcodes. These RNA barcodes are used to infer long-range projections of neurons. We have previously used BARseq to identify the projection patterns of different neuronal types defined either by gene expression [23,27] and/or their locations [22,28], and to identify genes that are associated with differences in projections within a neuronal populations [23]. Importantly, we showed that BARseq can resolve transcriptomically defined cell types of cortical neurons at cellular resolution by sequencing dozens of cell type markers [23]. Because BARseq has high throughput and low cost compared to many other spatial techniques [25,29-34], it is ideally suited for studying the spatial organization of gene expression at cellular resolution over whole brain structures, such as the cortex.

Here we use BARseq as a standalone technique for sequencing gene expression *in situ*, and scale it up to a brain-wide scale to resolve the distribution of neuronal populations and gene expression across the cortex. We generate a high-resolution map of 1.2 million cells with detailed gene expression. We find that although most neuronal populations are found in multiple cortical areas, the composition of neuronal populations is distinct across areas. By comparing area similarities in neuronal compositions, we uncover a modular organization of the cortex that matches cortical hierarchy and modules defined by connectivity in previous studies [4,5,11], thus revealing consistent areal specialization between the connectivity and molecular identity of cortical neurons.

Cellular resolution BARseq captures brain-wide gene expression in situ.

A number of recent single-cell transcriptomic studies [12,21,35-37] have classified neuronal types using hierarchical analyses, but these studies used different nomenclatures to refer to cell types at different levels of the hierarchy. To avoid confusion, we first define the cell type nomenclature we will use in this paper.

The highest hierarchical level we use, or H1 type, divides neurons into excitatory neurons, inhibitory neurons, and other cells; this level is referred to as "class" level in many studies. Within each H1 type, we subdivide neurons into H2 types, which are sometimes referred to as "subclasses"[12,13]. In particular, the cortical excitatory neurons can be largely divided into nine H2 types that are shared across most cortical areas, including L2/3 IT (intra-telencephalic neurons), L4/5 IT, L5 IT, L6 IT, L5 ET (extra-telencephalic neurons, also known as PT/pyramidal tract neurons), L6 CT (corticothalamic neurons), NP (near-projecting neurons), Car3, and L6b. This division follows recent single-cell RNAseq studies but differs from the classical tripartite of IT/PT/CT neurons, which were defined largely based on long-range projections and failed to capture differences among some transcriptomically distinct cell types, such as NP, Car3, and L6b. Each H2 type can be further divided into H3 types, which correspond to "cluster" or "type" level in some studies [12,13]. Previous studies have shown that H1 and H2 types are largely shared across most cortical areas, but the expression of many genes are localized to specific parts of the cortex both during development [14,38] and in the adult [39]. Clusters at the H3 level appeared to be enriched in neurons dissected from different parts of the cortex in single-cell RNAseq studies [12,21]. However, because these dissections could only be performed at a coarse resolution, the detailed distribution of neuronal populations at this granularity across cortical areas remains unclear.

To assess the distribution of neuronal populations at all three hierarchical levels across the cortex, we used BARseq to interrogate the expression of 107 cell type marker genes (Supplementary Table 1) in 40 hemi-brain coronal sections that cover the whole forebrain (Fig. 1A, B). We applied the same approach that we previously used to resolve cortical excitatory neuron types in the motor cortex [23]. Briefly, we selected marker genes that were optimized for distinguishing excitatory neuronal types in the cortex (see Supplementary Note 1). We evaluated this gene panel using a comprehensive single-cell RNAseq dataset from the cortex[13], and found that these genes performed similarly to the full transcriptome in distinguishing H2 (i.e. subclasses) and H3 (i.e. clusters) types (Fig. 1C; ED Fig. 1; see Supplementary Note 1). We used up to 12 padlock probes to target each gene, and each probe carried a 7-nt gene identification index (GII) that uniquely identified the gene. These GIIs were designed to have a minimum hamming distance of 3-nt to allow for error correction. We further included 5 blank GIIs that were not present in the padlock probes as negative controls when decoding the GIIs; these blank GIIs allowed us to control and evaluate false detection rates. We decoded the GIIs from seven rounds of sequencing using Bardensr [40] while maintaining an optimal false detection rate (~5% in this dataset). We registered our data to the Allen mouse brain common coordinate framework v3 (CCF v3)[41] using a semi-manual procedure that utilized QuickNii, Visualign [42], and custom python scripts (see Data and code availability). Three highly expressed high-level marker genes (*Slc17a7* for excitatory neurons, *Gad1* for inhibitory neurons, and *Slc30a3* for IT neurons) were detected by hybridization rather than by sequencing (Fig. 1B). We segmented cell bodies using DAPI as the nucleic signal and sequencing signals from all imaging channels as the cytoplasmic signal using Cellpose [43] (Fig. 1B), resulting in 1.2 million cells after quality control (QC, see Methods) with a mean of 60 unique reads/cell and 27 genes/cell (Fig. 1D, E).

At a gross anatomical level, many genes were differentially expressed across major brain structures and cortical layers (Fig. 1A). These expression patterns were consistent with the patterns of in situ hybridization in the Allen Brain Atlas [39](Fig. 1F). For example, classical cortical layer-specific markers, including *Cux2*, *Fezf2*, and *Foxp2*, were expressed in layer 2/3, layer 5/layer 6, and layer 6, respectively. *Rorb*, a layer 4 marker, was seen in layer 4 throughout the cortex except in the motor cortex and medial areas, which lack a classically defined layer 4. *Scnn1a* was expressed in primary visual areas and the retrosplenial cortex, but not other cortical areas. Thus, our dataset recapitulated known spatial distribution of gene expression.

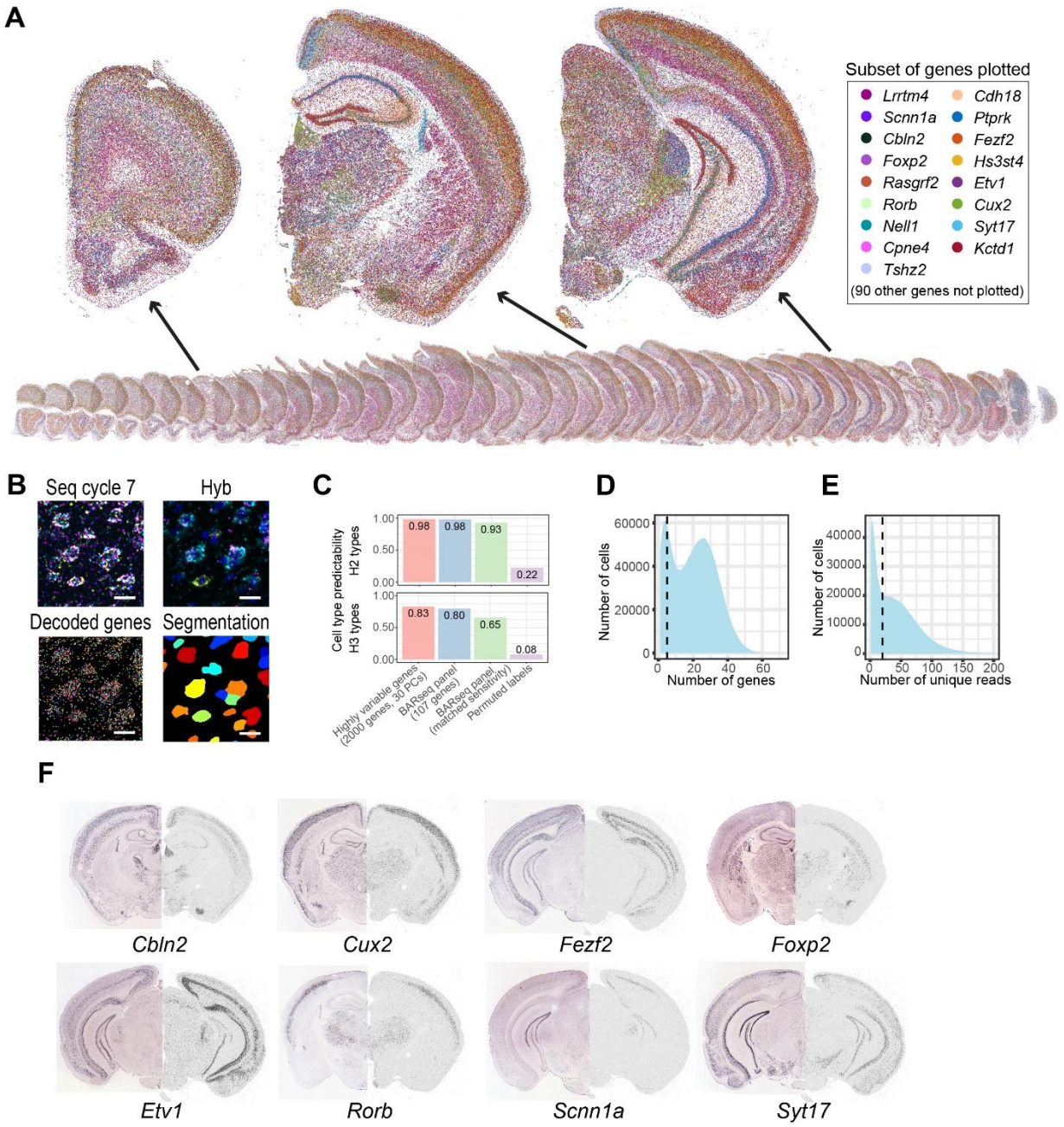


Figure 1. Interrogating brain-wide gene expression using BARseq.

*(**A**) Images of all 40 slices (bottom) and close-up images of three representative slices (top). For clarity, only 17 out of 107 genes (indicated on the right) are plotted. (**B**) Close up images of the last sequencing cycle, hybridization cycle, decoded genes, and cell segmentation. Scale bars = 10 μm. (**C**) Single-cell cluster assignment performance using the full transcriptome, top principal components, and the 107-gene panel for H2 (top) and H3 (bottom) clusters. (**D**) Gene counts per cell and (**E**) read counts per cell in the dataset. Quality control thresholds are indicated by dashed lines in both plots. The lower peaks in gene and read counts likely include non-neuronal cells that do not express the cortical neuronal markers in our gene panel and non-cellular particles that are fluorescent. (**F**) The expression patterns of representative genes in Allen Brain Atlas (left half) compared to the current dataset (right half).*

De novo clustering reveals neuronal subpopulations that are consistent with reference transcriptomic types

We next identified transcriptomic types of individual neurons based on single-cell gene expression. Generally, two approaches can be used to identify cell types in new transcriptomic datasets. In the first approach, we can map individual neurons in a new dataset directly to clusters in reference single-cell RNAseq datasets to determine cell type identities. This approach can match small datasets to cell types discovered in much larger, higher-resolution datasets [44-46], but is prone to technique-specific variations when mapping data generated by different techniques [47]. Alternatively, we can cluster the new dataset and map *clusters* to cell types in reference single-cell RNAseq datasets [12,48]. This approach can better account for technique-specific variations and batch effects [49], but the ability to distinguish cell types is dependent on both data quality and sample size in the new dataset. Because our dataset contains 1.2 million cells, which is comparable in size to the most comprehensive single-cell RNAseq datasets generated to date [13,36,50], we reasoned that *de novo* clustering followed by assessment at the cluster level would be more easily interpretable.

We applied hierarchical clustering to separate neuronal populations at H1, H2, and H3 levels (Fig. 2A; see Methods). Clustering all cells resulted in 24 clusters, which we then combined into three H1 types (642,340 excitatory neurons, 427,939 inhibitory neurons, and 188,977 other cells) based on the expression of *Slc17a7* and *Gad1* (ED Fig. 2A). Of these 1.2 million cells, 517,428 were in the cortex and were the focus of the remainder of this study. Previous studies estimated the fraction of inhibitory neurons in the mouse cortex to be between 10% and 20% [51,52]. Consistent with these estimates, 16% of neurons in the cortex were inhibitory neurons in our dataset (427,766 excitatory neurons, 83,394 inhibitory neurons, and 6,268 other cells). Because we only sampled *Slc17a7* and *Gad1* for excitatory and inhibitory neuron markers, the excitatory neurons identified were dominated by cortical neurons, although we also saw neurons in the pons and the epithalamus in this group. The third group of cells, which expressed neither *Slc17a7* nor *Gad1*, included subpopulations of subcortical neurons (e.g. the midbrain and the thalamus) and non-neuronal cells (e.g. glial cells, immune cells, and epithelial cells). We expect this group of cells to be undersampled, because these cells may not express cortical cell type marker genes we probed at sufficient levels to pass quality control (Fig. 1D, E). Based on the fraction of excitatory neurons that express both *Slc17a7* and *Gad1*, we estimated that the probability of segmentation errors in which two neighboring cells were merged together, i.e. doublet rate, to be between 5% and 7% (ED Fig. 2B, C; see Supplementary Note 2). The 24 clusters, comprising the three H1 types, largely corresponded to coarse anatomical structures in the brain (Fig. 2B). For example, different clusters were enriched in the lateral group and ventral group of the thalamus, the intralaminar nuclei, the epithalamus, the medial, basolateral, and lateral nuclei of the amygdala, the striatum, and the globus pallidus (Fig. 2B). These results recapitulated the clear distinction of transcriptomic types across anatomically defined brain structures as observed in whole-brain single-cell RNAseq studies [36,50](Zeng, pers. comm.)

We then re-clustered the excitatory and inhibitory neurons separately into H2 types (Fig. 2C; ED Fig. 2D) to improve the resolution of clustering. At this level, we recovered major inhibitory neuron subclasses (Pvalb, Sst, Vip/Sncg, Meis2-like, and Lamp5) and all cortical excitatory subclasses (L2/3 IT, L4/5 IT, L5 IT, L6 IT, L5 ET, L6 CT, NP, Car3, L6b) observed in previous cortical single-cell RNAseq datasets [12,13,21,53]. The H2 types expressed known cell type markers and other highly differentially expressed genes (Fig. 2D). For example, *Cux2* is expressed mostly in superficial layer IT neurons and Car3 neurons; *Fezf2* is expressed in NP and L5 ET neurons; and *Foxp2* is expressed specifically in L6 CT neurons (see Supplementary Note 3 for detailed description of all genes). Although we generated the full 40-section data in two batches (see

Methods), we did not observe strong batch effects, because neurons from different slices across the two batches were intermingled in the UMAP plot for excitatory H2 types (ED Fig. 2E). Thus, the H2 types recapitulated known neuronal types at medium granularity that were identified in previous single-cell RNAseq datasets [12,13,21].

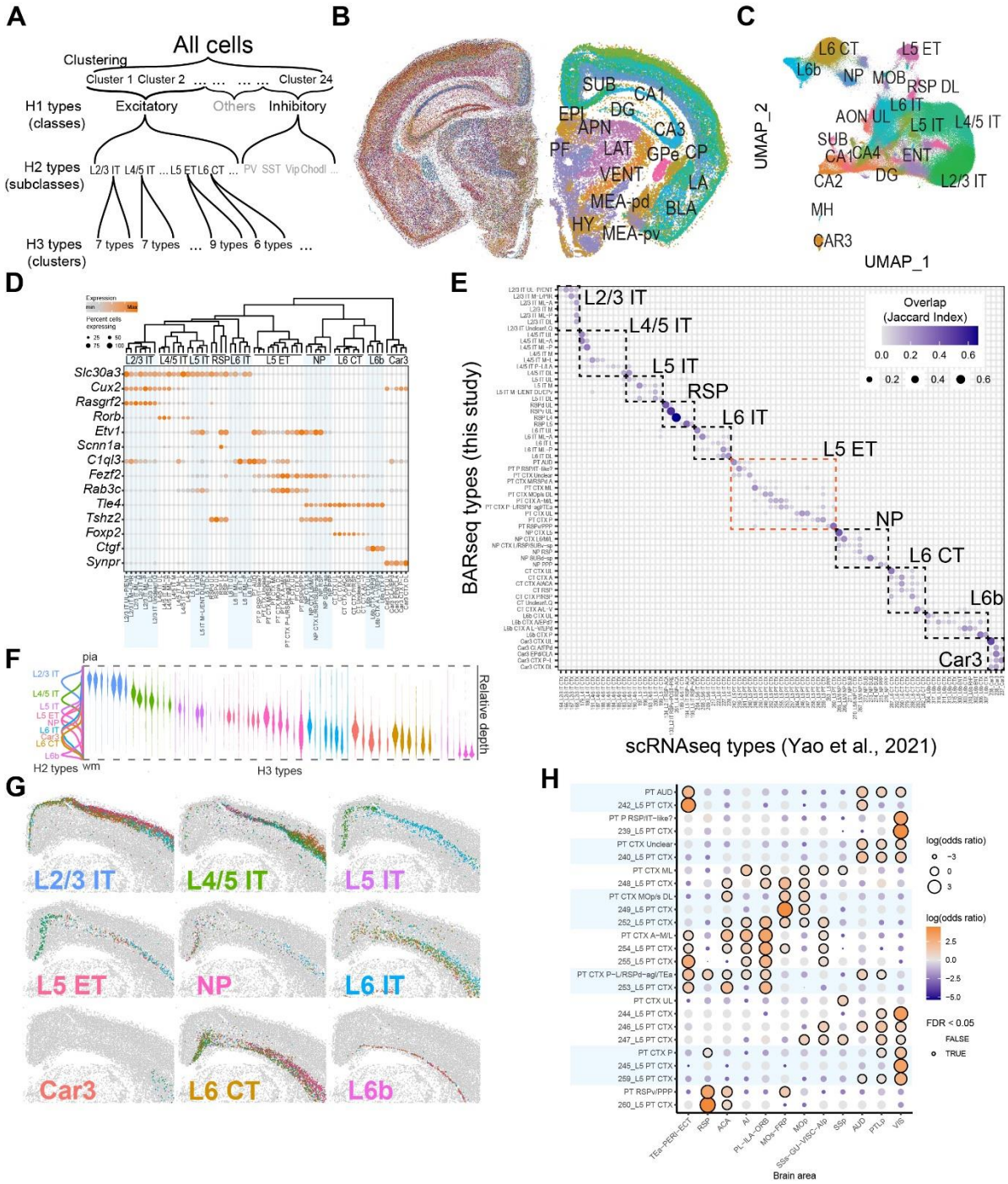


Figure 2. BARseq captures gene expression and spatial distribution of cortical excitatory cell types.

*(**A**) Workflow of hierarchical clustering. (**B**) Gene expression (left) and H1 clusters (right) in a*

*representative slice. Major anatomical divisions that are resolved by H1 clusters are labeled. SUB: subiculum, DG: dentate gyrus, CP: caudate putamen, GPe: Globus Pallidus, external segment, LA: lateral amygdala, BLA: basolateral amygdala, MEA-pd(pv): medial amygdalar nucleus, posterodorsal (posteroventral), EPI: epithalamus, APN: anterior pretectal nucleus, LAT: lateral group of the dorsal thalamus, VENT: ventral group of the dorsal thalamus, PF: parafascicular nucleus, HY: hypothalamus. (**C**) UMAP plot of gene expression of excitatory neurons colored by H2 types. (**D**) Marker gene expression in H2 types. Colors indicate mean expression level and dot size indicates fraction of cells expressing the gene. The dendrogram on top shows hierarchical clustering of pooled gene expression within each H2 type. (**E**) Overlap between BARseq H3 types and scRNAseq cell types from Yao, et al. [13]. Dashed boxes indicate the parent H2 types. (**F**) The laminar distribution of H2 types (shown on the left) and each H3 type. H3 types are sorted by their median laminar position. Colors match those used in (C). (**G**) The distribution of H3 types in the dorsomedial portion of the cortex on a representative slice. The parent H2 types are indicated in each plot. (**H**) Distribution across CCF regions of matching L5 ET BARseq H3 types and L5 ET scRNAseq cell types (Jaccard Index > 0.1). Matched types are shown next to each other and share the same background color).*

We then re-clustered each excitatory H2 type into H3 types. To quantify how well H3 types corresponded to reference transcriptomic types identified in previous single-cell RNAseq studies, we used a k-nearest neighbor-based approach to match each H3 type to leaf-level clusters in a previous single-cell RNAseq dataset [13] (see Methods). We found that cortical H2 types had a one-to-one correspondence to subclass-level cell types in the single-cell RNAseq data (Fig. 2E). Within each H2 type, the H3 types differentially mapped onto single or small subsets of leaf-level clusters in the single-cell RNAseq data (Fig. 2E; see ED Fig. 2F for matching of clusters outside of the cortex). These results demonstrate that our dataset resolved fine-grained neuronal subpopulatoins corresponding to clusters obtained previously using single-cell RNAseq data.

Both H2 types and H3 types were organized in an orderly fashion along the depth of the cortex. H2 types were concentrated in distinct layers, whereas H3 types were concentrated in sublayers, or finer divisions within each layer (Fig. 2F). For example, multiple H3 types of L2/3 IT, L4/5 IT, and L6 IT clearly occupied distinct sublayers in the somatosensory cortex area (Fig. 2G). These results are consistent with previous studies using other spatial transcriptomic techniques [30,48,54]. Thus, our data captured the laminar organization of cortical excitatory neurons.

Previous single-cell RNAseq studies, which used manual dissection to distinguish neurons from different cortical areas, showed that subpopulations of cortical neurons were also differentially distributed across large areas of the cortex [13,21]. Consistent with these studies, H3 types and their matching clusters in single-cell RNAseq datasets were also found in similar cortical areas (Fig. 2H, ED Fig. 2G-H, ED Fig. 3). For example, the H3 type "PT AUD" and its corresponding single-cell RNAseq cluster (242_L5_PT CTX) were both enriched in the lateral cortical areas (TEa-PERI-ECT) and auditory cortex (AUD), whereas the H3 type "PT CTX P" and its corresponding single-cell RNAseq clusters (245_L5_PT CTX and 259_L5_PT CTX) were all enriched in the visual cortex (VIS). Thus, at this coarse spatial resolution, our data recapitulated previously observed areal distribution of cortical excitatory neurons.

To summarize, our data resolved fine-grained transcriptomic types of cortical excitatory neurons that were consistent with previous single-cell RNAseq datasets [13] and recapitulated their areal and laminar distribution [13,21,48]. Unlike these previous studies, the high resolution and the cortex-wide span of our dataset allow us to go beyond this coarse spatial resolution and resolve the spatial enrichment of gene expression and the distribution of neuronal subpopulations across the cortex at micron-level resolution. This high resolution matches those in previous connectivity studies [4,5], and thus enables comparison of transcriptomic

and connectivity organization of the cortex. In the next sections, we first examine how gene expression varies across the tangential plane of the cortex. We then explore the distribution of neuronal populations across cortical areas. Finally, we identify modules of cortical areas based on the composition of neuronal subpopulations.

Gene expression varies within H2 types along interconnected cortical areas.

Gene expression varies substantially across the whole cortex [39,55], but most cortical areas largely share the same H2 types, or subclasses, of excitatory neurons [13,21]. Thus, it is unclear how differences in the organization of neuronal subpopulations lead to area-specific gene expression. Three sources of variation could contribute to the differences in gene expression across areas (Fig. 3A). First, the composition of H2 types may drive the differences in gene expression across the cortex (Fig. 3Aa, the cell type composition model). For example, the ratio of H2 type X to type Y might be high in visual cortex but low in motor cortex, so genes that are expressed more highly in X than in Y will be more highly expressed in visual cortex. Second, the expression of some genes may vary across space regardless of H2 type, i.e. they change consistently across space in multiple H2 types (Fig. 3Ab, the spatial gradient model). In this model, gene A may be more highly expressed in the visual cortex than in the motor cortex in both H2 types X and Y. Finally, the expression of some genes may vary across space in an H2-specific manner (Fig. 3Ac, the area-specialized cell type model). For example, gene A may be more highly expressed in the visual cortex than in the motor cortex in H2 type X. Our data suggest that all three models contribute to the spatial variation of gene expression in the cortex, but the model that contributes most to the variation for each gene can vary.

To determine the contribution of each source to the variation of gene expression across areas, we discretized the cortex on each coronal slice into 20 spatial bins. Each bin spanned all cortical layers, and bin widths were chosen so that each bin had the same number of cells (see Methods, ED Fig. 4A). We then assessed how much the variation in bulk gene expression across bins can be explained by space or by composition of H2 or H3 types using one-way ANOVA (Fig. 3B, C; see Methods). We found that variations in many genes were strongly explained by the composition of H2 types; these patterns were consistent with the cell type composition model (Fig. 3Aa). An example gene of this category is *Ctgf*, which is specifically expressed in L6b neurons at a consistent level across space (Fig. 3D, *top*). Thus, variations in the expression (up to 80%) of *Ctgf* across space were largely explained by the fraction of L6b neurons in each bin rather than by variation in gene expression within cells. Other genes, in contrast, were largely explained by the composition of H3 types rather than the composition of H2 types (Fig. 3C, i.e. the area-specialized cell type model in Fig. 3Ac). Many genes in this category were also highly spatially variable (Fig. 3C, rho = 0.23), which suggest that H3 types were likely differentially distributed across the cortex (e.g. *Nnat*, marker of lateral areas, Fig. 3D, *middle*). Finally, some genes displayed high spatial variability, but relatively low H2 and H3 variability. These genes were usually expressed in multiple H2 types (e.g. *Tenm3*, Fig. 3D, *bottom*) and varied consistently in space across these H2 types, suggesting a general spatial gradient that is independent of H2 types. The spatial patterns of these genes were consistent with the spatial gradient model (Fig. 3Ab).

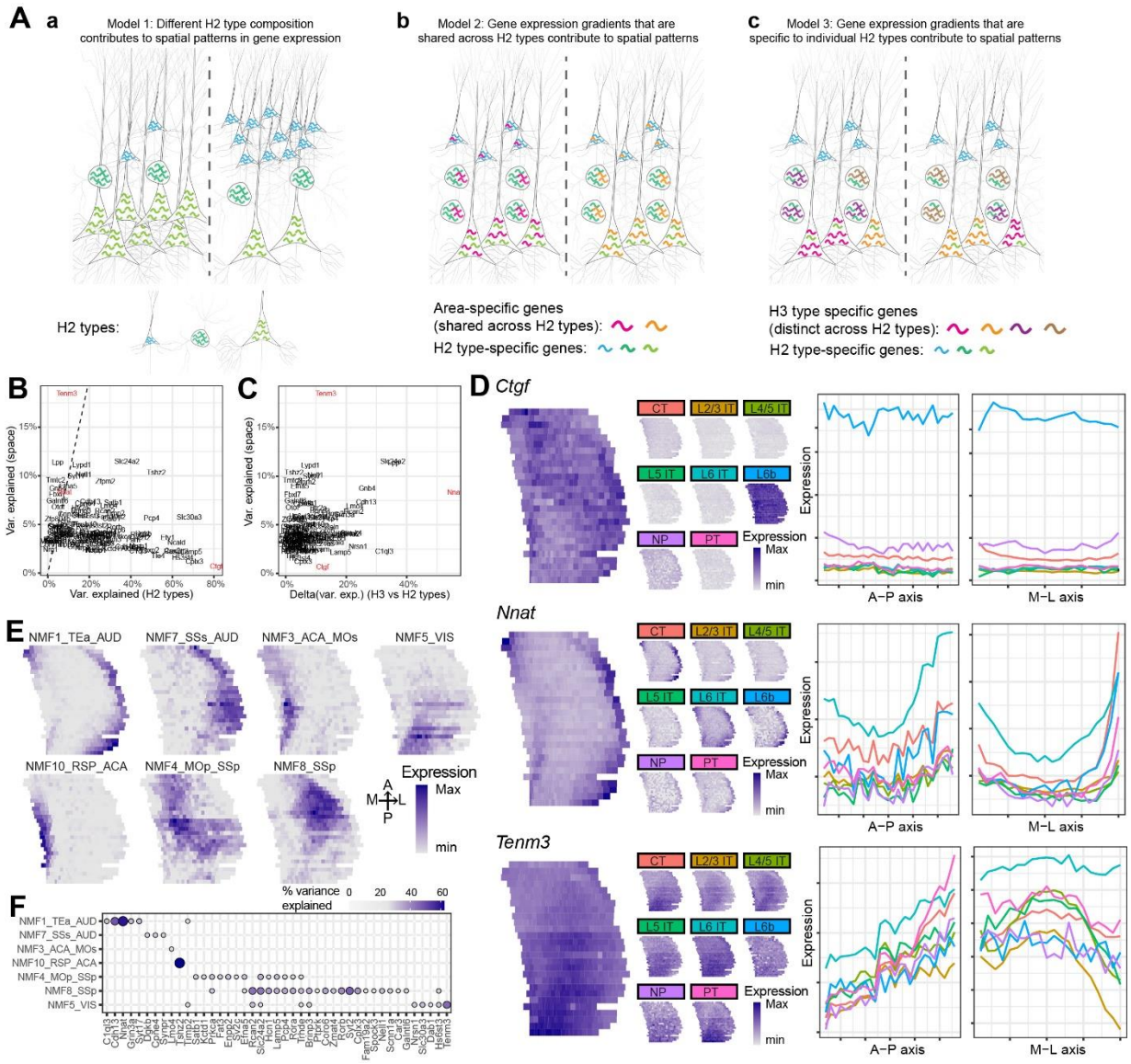


Figure 3. Spatial variations of gene expression across the cortex.

*(**A**)Three models of differential gene expression across cortical areas. (a) The relative proportion of each H2 type is different across areas. (b) The expression of genes varies across cortical areas consistently in different H2 types. (c) In each area, H2 types are enriched in different area-specialized H3 types. Different H2 types are indicated by neurons of different shapes. Cool-toned squiggles indicate H2 type-specific genes. Warm-toned squiggles indicate area-specific genes expression in (b) and H3 type-specific genes in (c). (**B**)(**C**) Variance in gene expression explained by space compared to that explained by H2 types (**B**) or additional variance explained by H3 types (**C**). (**D**) The expression pattern of the indicated genes plotted on a flat map of the cortex in all cells (left) or in each H2 type (center). The variations of gene expression in each H2 type along the AP axis and the ML axis are shown on the right. Line colors indicate H2 types as shown in the center plots. (**E**) The expression of selected spatially variant NMF modules plotted on a cortical flatmap. (**F**) The expression of select marker genes in each NMF module.*

Because the spatial patterns of many genes were similar across genes and H2 types, we sought to extract basic spatial components that were shared across genes and H2 types using non-negative matrix factorization (NMF) [56]. Briefly, we performed NMF on the residuals of gene expression of spatial bins after accounting for differences in the composition of H2 types in each bin. We extracted ten NMF components, seven of which captured spatial variations in gene expression (the other three components captured slice-specific technical variability and were not used for subsequent analyses; see Supplementary Note 4 and ED Fig. 4B, C). We found that the majority of NMF components were expressed not along broad spatial gradients along major axes, but rather in areas that were functionally related and highly interconnected (Fig. 3E; ED Fig. 4D). For example, NMF5 was expressed mostly in the visual areas, whereas NMF8 was expressed in somatosensory areas. Other NMF modules, including NMF1 (medial areas) and NMF10 (lateral areas), were expressed in combinations of areas that were functionally distinct but also highly interconnected [4,5]. Individual spatially variant genes were usually strongly associated with only one or two components (Fig. 3F; ED Fig. 4E; ED Fig. 5), and the association recapitulated known spatial patterns of these genes. For example, *Tenm3* was expressed mostly in posterior sensory areas, including the visual cortex, auditory cortex, and part of the somatosensory cortex[39](Fig. 3D, *bottom*); *Tenm3* was strongly associated with NMF5 (Fig. 3F), which was also expressed in the same sets of areas (Fig. 3E). Thus, the finding that gene expression varies along sets of interconnected areas suggests an intriguing link between gene expression and intra-cortical connectivity across areas.

Cortical areas have distinct compositional profiles of H3 types.

Because the spatially varying NMF modules were obtained after controlling for variability in the composition of H2 types, but not H3 types, we hypothesized that these modules reflected differences in the composition of H3 types across cortical areas. Consistent with this hypothesis, each H3 type was enriched in a small subset of NMF modules (Fig. 4A; see Methods). All H2 types contained at least one H3 type that was associated with NMF modules that were expressed in the medial and lateral areas (NMF 1, 3, 7, 10), and one to four H3 types that were associated with NMF modules that were expressed in subsets of the dorsal cortex, including the motor, somatosensory, and visual areas (NMF 4, 5, 8). The associations between H3 types and the NMF modules were different across H2 types, suggesting that different H2 types were specialized to different degrees at the H3 level. For example, the H3 types of L5 ET neurons had strong associations with individual NMFs, whereas H3 types of NP and L6b neurons showed little specialization within the dorsal cortex.

Consistent with the association in gene expression between NMF modules and H3 types, the H3 types also overlapped with their corresponding NMF modules in space (Fig. 4B; ED Fig. 6; see Methods). For example, L4/5 IT ML-P was associated with NMF5_VIS and was enriched most strongly in visual cortex; similarly, L4/5 IT P-L/LA was associated with NMF1_TEa_AUD and NMF7_SSs_AUD and was most highly enriched in auditory cortex and temporal association areas. Consistent with the expression patterns of NMF modules, many of these sets of areas were largely within cortical modules that were defined by inter-connectivity [4,5]. Thus, H3 types are associated with spatial gene co-expression modules and, at a coarse spatial resolution, are enriched in combinations of cortical areas that are highly interconnected.

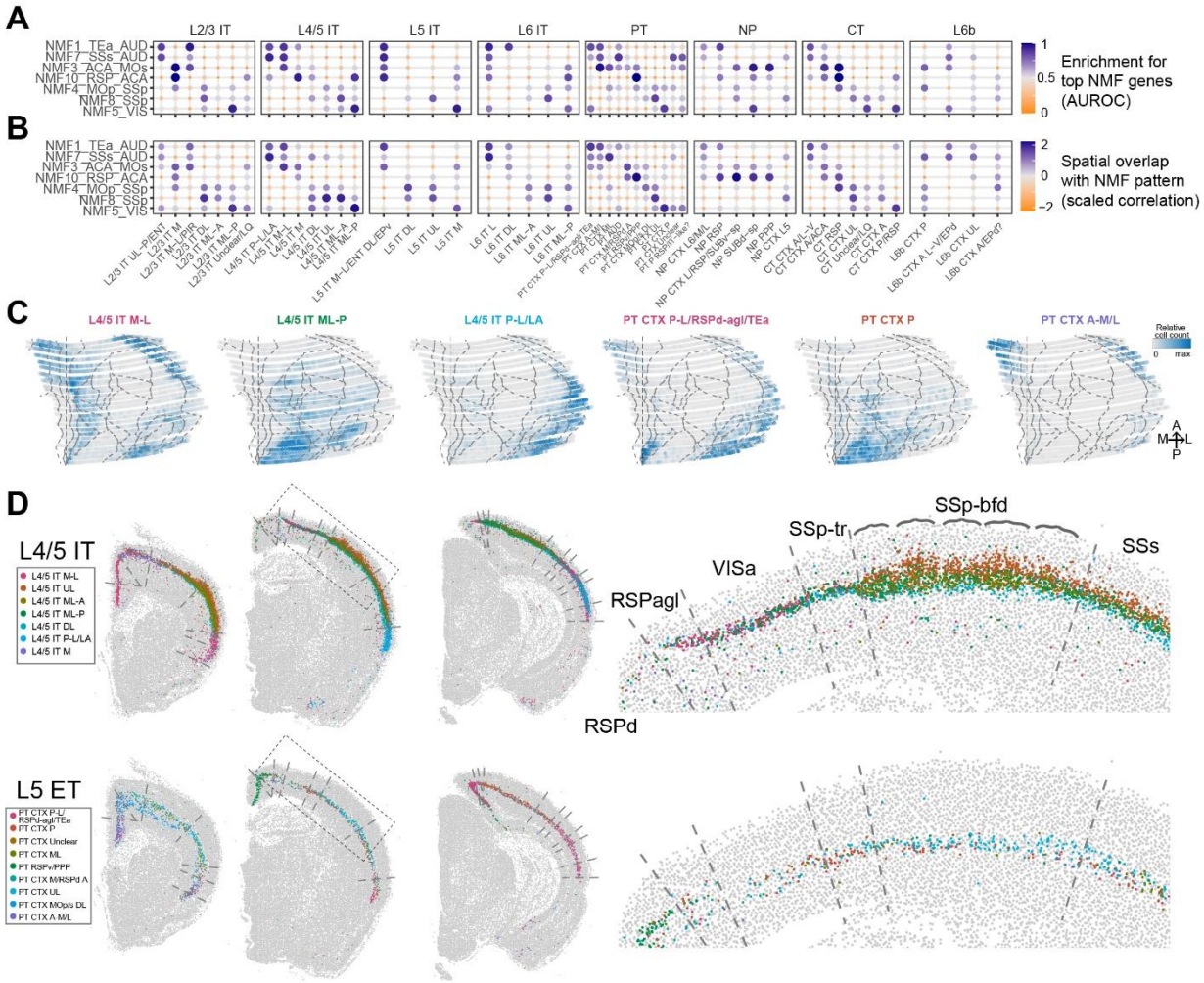


Figure 4. The composition of cell types is distinct across cortical areas.

*(**A**) AUROC of the enrichment for top NMF genes in each H3 type (see Methods). (**B**) Overlap between the spatial patterns of NMF module expression and the spatial distribution of H3 types. (**C**) The spatial distribution of example L4/5 IT and L5 ET H3 types across the cortex. Intensities indicate relative cell counts in each cubelet. Dashed lines delineate cortical areas. (**D**) The distribution of L4/5 IT and L5 ET H3 types in example coronal sections. Dashed lines indicate area borders in CCF. Magnified views of the dashed boxes are shown on the right. Brackets in the top plot indicate barrels in the barrel cortex.*

To further assess the areal distribution of H3 types, we discretized the cortex on each coronal slice into "cubelets" with similar widths along the mediolateral axis across all slices. Each cubelet spanned all cortical layers and was about 100 μm to 200 μm wide along the curvature of the cortex, and 20 μm thick along the A-P axis (i.e. the thickness of each coronal section; see Methods; ED Fig. 4A). Thus, unlike the "spatial bins," which contained similar number of neurons within a slice, "cubelets" were of similar physical sizes across all slices. We found that most H3 types were shared by multiple cortical areas and not specific to any single area (Fig. 4C; Supp. Fig. 1). Thus, the distinctness of neighboring cortical areas cannot be explained simply by the presence or absence of an area-specific H3 type. However, we noticed that the compositional profiles of H3 types often changed abruptly near area borders defined in CCF (Fig. 4D). Most salient changes occurred at the lateral and medial areas, which are consistent with single-cell RNAseq data [13]. Within the dorsolateral cortex, although neighboring cortical areas sometimes shared sets of H3

types, their compositions usually changed abruptly at or near area borders. For example, three L4/5 IT types, including L4/5 IT UL, L4/5 IT ML-P, and L4/5 IT DL, were found in three adjacent somatosensory areas (trunk area, SSp-tr; barrel cortex, SSp-bfd; secondary somatosensory cortex, SSs; Fig. 4D, *top row*), but the numbers of neurons of each type were distinct across the three areas. L4/5 IT UL was found in only small numbers in SSp-tr, but expanded in both the number of neurons and their laminar span in the barrel cortex. Furthermore, L4/5 IT UL neurons were clustered along the mediolateral axis into structures that resembled barrels. In the secondary somatosensory cortex, the distribution of L4/5 IT UL lost the barrel-like patterns but remained present in substantial numbers. In contrast, L4/5 IT DL became more dominant relative to L4/5 IT ML-P. Similarly, in L5 ET neurons, PT CTX P was more dominant in SSp-tr, whereas PT CTX ML was found mostly in SSs (Fig. 4D, bottom row). To quantify how well abrupt changes in H3 type composition corresponded to borders between all cortical areas, we identified positions where H3 type composition changed abruptly by identifying peaks in the absolute value of the first derivatives of H3 type composition along the ML axis within each slice. Consistent with the impression from images of coronal sections, 53% of peaks in the first derivatives were within 150 µm from the closest CCF border (we could not assess matching using a more stringent distance because 150 µm is already comparable to the cubelet size in this dataset). The fraction of peaks that were close to CCF borders was higher than 99% of shuffling controls (ED Fig. 7A). These results reconcile two seemingly contradictory observations in previous single-cell RNAseq studies: distant cortical areas (e.g. visual cortex and anterolateral motor cortex) have distinct excitatory cell types [21], but individual cell types are usually found across large areas of the cortex [13]. Thus, cortical areas are largely distinct in their compositional profiles of H3 types, but individual H3 types are rarely specific to a single cortical area.

H3 type composition reveals modular organization of the cortex.

Because variation in gene expression and H3 types both followed spatial patterns that were reminiscent of highly interconnected cortical areas, we hypothesized that, like the modular organization of corticocortical connectivity, cortical areas were also organized into modules based on H3 types: cortical areas within a module would be composed of similar sets of H3 types, whereas areas in different modules would be more distinct in their compositions of H3 types. To test this hypothesis, we first asked how well cortical areas could be distinguished using the compositions of H3 types. We then identified cortical modules using the differences in H3 type compositions.

To test how well the compositions of H3 types could predict cortical areas, we first used random forest classifiers to predict the AP and ML coordinates of each cubelet given either the total gene expression in that cubelet (Fig. 5A) or its H3 type composition (i.e. the fraction of each H3 type within a cubelet; Fig. 5B). We found that cubelet gene expression was highly predictive of locations in the cortex, capturing 94% of variance on both the AP and ML axes. The distance between the predicted and true location of a cubelet across the whole cortex was 235 ± 270 µm (median ± std) along the AP axis (spanning 5,900 µm) and 245 ± 364 µm along the ML axis (spanning 8,400 µm) (Fig. 5A, C). These prediction errors were close to the sampling frequency in our dataset (200 µm between adjacent slices on the AP axis, and 100 µm to 200 µm cubelet width on the ML axis). Strikingly, the H3 type compositions performed similarly well, capturing 89% variance on the AP axis and 92% variance on the ML axis (the prediction errors were 312 ± 360 µm on the AP axis and 269 ± 402 µm on the ML axis, median ± std; Fig. 5B, C). Consistent with the high precision in predicting the absolute locations in the cortex, both gene expression and the composition of H3 types were highly predictive of the area labels in CCF (75% correct using gene expression and 69% correct using H3 type composition, compared to 8% in shuffled control; Fig. 5D, E). H3 types within a single parent

H2 type were also somewhat predictive of cubelet locations, but those of H2 types in superficial layers (e.g. L2/3 IT and L4/5 IT) were generally more predictive of cubelet locations along the ML axis than those of H2 types in the deep layers (e.g. L6 IT and L6 CT) (ED Fig. 7B, C). The predicted maps correctly captured the locations of most cortical areas, and most of the incorrect predictions occurred along the borders of areas (Fig. 5D). Thus, both cubelet gene expression and H3 type compositions are highly predictive of the locations along the tangential plane of the cortex and the identity of the cortical areas.

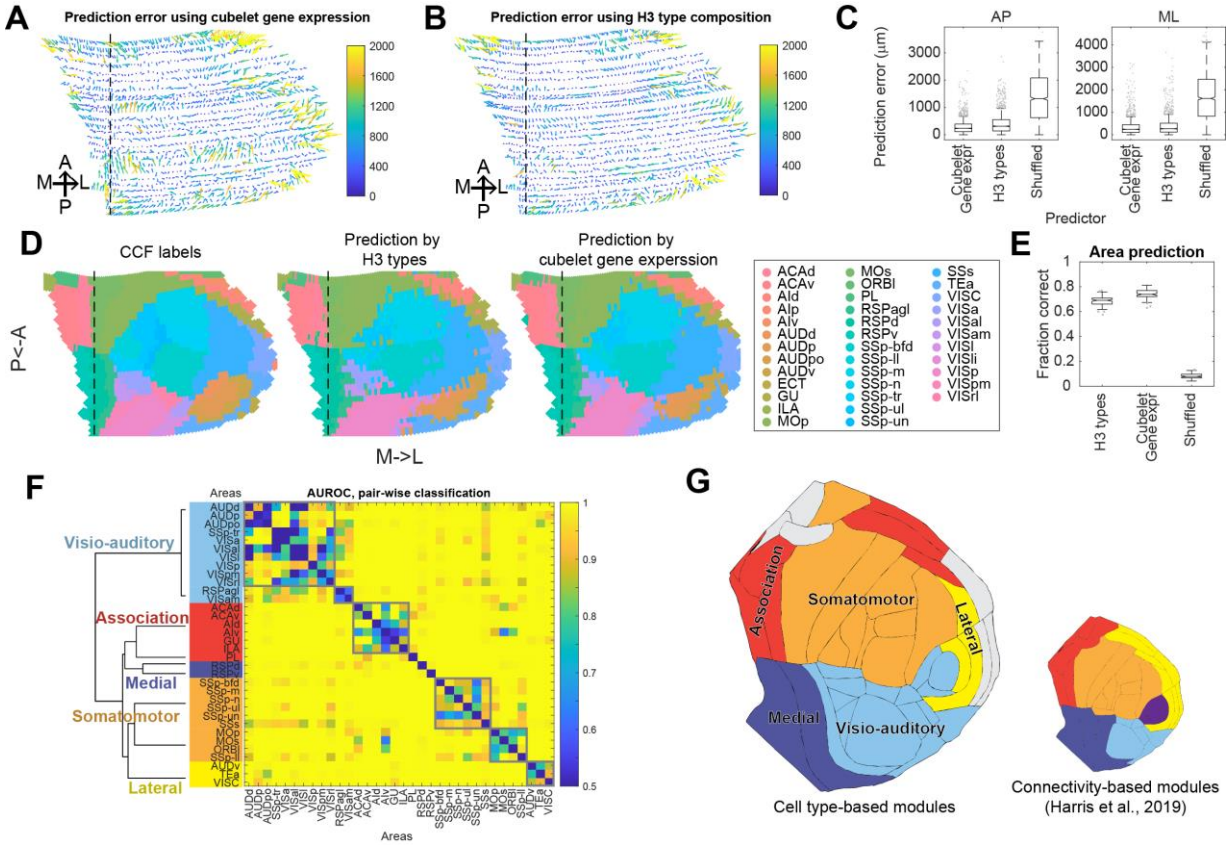


Figure 5. Modular and hierarchical organization of cortical areas by cell types.

*(**A**)(**B**) Heatmaps showing the errors in predicting cubelet locations using gene expression (**A**) or H3 type composition (**B**). Arrows indicate the directions of the errors and colors indicate the magnitudes of the prediction errors (in μm). The lengths of the arrows are proportional to the prediction error. (**C**) Box plots summarizing the prediction performance shown in (A) and (B). Boxes show median and quartiles and whiskers indicate range after excluding outliers. Dots indicate outliers. (**D**) Cortical areas defined in CCF (left) and those predicted by H3 types (center) and cubelet gene expression (right). (**E**) Fraction of correctly predicted cubelets using H3 type composition, cubelet gene expression, and shuffled control. (**F**) Similarity matrix of cortical areas. Areas are sorted by modules, which are color-coded on the left. Dendrogram is calculated using similarity of H3 type composition. Clusters obtained based on the similarity matrix are shown in gray boxes. (**G**) Cortical flatmaps colored by cell type-based modules as defined in (F) (left) and by connectivity-based modules identified by Harris, et al. [4] (right). Areas in gray did not contain sufficient numbers of cubelets and were excluded from this analysis.*

We next assessed the similarity and modularity of cortical areas based on how well cortical areas could be distinguished by their H3 type compositions (Fig. 5F; see Methods). Briefly, we built a distance matrix between cortical areas based on how well they can be distinguished pairwise using H3 type composition, then performed Louvain clustering on the distance matrix. We identified six clusters, each of which

consisted of more than one area (Fig. 5F, gray boxes); these included two clusters that corresponded to the visio-auditory areas and one cluster each for the association areas, somatosensory cortex, motor cortex, and lateral areas. This modular organization is robust to small errors in CCF registration (ED Fig. 7D; see Methods). We further combined these clusters with areas that did not cluster with other areas (PL, RSPd, RSPv) into cortical modules based on similarity in H3 type composition. These modules largely included the visio-auditory areas, somatomotor areas, the association areas, medial areas, lateral areas, respectively (Fig. 5F, G). Strikingly, these modules were largely consistent with cortical modules that are highly connected (connectivity-based modules) [4]. Thus, highly interconnected cortical areas also share similar H3 types and, consequently, characteristic transcriptomic signatures. This relationship between compositions of H3 types and connectivity reflects a novel feature of the mesoscale connectivity between cortical areas that is distinct from the known stereotypical connectivity of cell types in a canonical cortical microcircuit.

Discussion

Using BARseq, we have generated a cortex-wide map of transcriptomic types of excitatory neurons with high transcriptomic and spatial resolution. Our map revealed that the spatial patterns of gene expression and the compositions of neuronal types delineate the divisions of cortical areas. This area-based specialization in gene expression further reveals similarity across subsets of cortical areas that are also highly interconnected. Together, these results reveal a modular organization of the cortex that is consistent across transcriptomically defined neuronal types and connectivity.

Fine-grained neuronal types reflect cortical area-based specialization

By building a high-resolution map of gene expression across the cortex, our data reveals the transcriptomically defined neuronal type and gene expression basis of cortical areas. This map of transcriptomically defined neuronal types can serve as an "anchor" to associate other neuronal properties and provides a foundation for understanding the structural and functional specialization of cortical areas, which are historically defined by cytoarchitecture [1-3]. Because it is easier to associate neuronal morphology, electrophysiology, connectivity, and neuronal activity with gene expression than with cytoarchitecture, redefining cortical areas and potentially other anatomical divisions in the brain using transcriptomic type information could produce a next-generation reference atlas that is more informative than the existing cytoarchitecture-based atlases. Because of individual differences across brains, building such a transcriptomics-based reference atlas would ideally involve averaging across thousands of individual brains [41], all interrogated at high spatial and transcriptomic resolution. BARseq is ideally suited for such an approach, because it is extremely high-throughput and low-cost compared to other spatial transcriptomic approaches. For example, we collected the hemisphere in this study in seven days of sequencing using a single microscope and spent $3,000 for reagents. Our study serves as a proof-of-principle for this approach, towards building a transcriptomic type-based reference atlas.

Transcriptomically defined neuronal types are largely hierarchical [16], yet the biological basis for such hierarchy remains unclear. For example, cortical excitatory neurons and inhibitory neurons are segregated at a coarse granularity in the transcriptomic taxonomy regardless of the cortical area they are from [13,21](ED Fig. 2A). In contrast, cortical areas have a strong influence on the activity of both excitatory and inhibitory neurons in response to sensory stimuli. Thus, the hierarchy of transcriptomic taxonomy does not necessarily reflect similarity in stimulus-evoked activity or other functional measures. What, then, generates the observed hierarchy of transcriptomically defined neuronal types across numerous transcriptomic studies?

Our results provide an interesting clue to this question. We found that the medium-grained H2 types represent general cell types that are shared across all cortical areas. In contrast, fine-grained H3 types are restricted in distribution and reflect area specialization. We speculate that different developmental processes established divisions at these two hierarchies: Divisions of H2 types are driven by developmental timing and other genetic mechanisms that are shared across radial glial cells [57], whereas divisions across H3 types are likely driven by mechanisms that contribute to cortical arealization [14,38,58]. These biological processes are not inherently hierarchical, but because they affect potentially different sets of genes at different stages of development, their "effects" on the overall gene expression observed in adult neurons differ in magnitude and thus *appear* hierarchical. Consistent with this hypothesis, the observation that spatial gradients in the expression of some genes were shared across H2 types also supports a non-hierarchical origin of the observed hierarchical organization of transcriptomically defined neuronal types. Comprehensive interrogation and analysis of gene expression during cortical development combined with lineage tracing will provide crucial evidence to support or refute our hypothesis.

Wire-by-similarity describes cortical organization at the mesoscale

Our results indicate that cortical areas that are highly interconnected also have similar H3 types. This relationship between transcriptomically defined neuronal types and connectivity, which we summarize as "wire-by-similarity," is distinct from conventional wiring rules that are commonly observed at individual neuronal type level (Fig. 6). In most circuits, similar cell types usually share similar inputs and/or outputs (Fig. 6A). For example, different subtypes of retinal ganglion cells all receive inputs from bipolar cells, and then project to the tectum and/or the thalamus[16]. Retinal ganglion cells, however, are not highly connected with each other, as in a wire-by-similarity organization (Fig. 6B). Similarly, in the cortex, neurons that belong to the same types are not necessarily highly connected (e.g. *Sst* neurons are sparsely connected with each other [59]). Wire-by-similarity is thus not a trivial consequence of cell type-specific connectivity observed at a cortex-wide scale, but it is also *not in conflict* with conventional cell type wiring rules. For example, in the model shown in Fig. 6, neurons of cell type A and A´ could follow a rule in which they project to areas with similar types; such a rule would obey both cell type-specific connectivity and wire-by-similarity. These examples illustrate the difference in scale at which wire-by-similarity and conventional cell type wiring rules apply: wire-by-similarity does not describe the connectivity of individual neuronal types, but how divisions within a large brain region (i.e. areas within the cortex) relate to each other in terms of cell types and connectivity.

This wire-by-similarity organization may provide an explanation for the striking lack of projection specificity among different transcriptomic types of corticocortical projection neurons [21,27,48]. Cortical projection neurons were historically defined by long-range projections. These projection-defined neuronal populations largely matched transcriptomic classifications at the level of IT, L5 ET, and L6 CT. Finer transcriptomic divisions within the IT neurons, which contribute to the majority of corticocortical projections, however, correspond poorly to differences in projection patterns [21,27,48]. Although our data do not provide a direct link between transcriptomic types and projections, the wire-by-similarity organization implies that in two areas with similar transcriptomic types, the same sets of IT neurons likely project to each other. Such connectivity may be difficult to specify if corticocortical projections is established largely by a static "zip code," such as a signaling gradient, because the neurons from the two areas would need to send projections along opposite directions. In contrast, this type of connectivity is compatible with temporally controlled genetic programming [60]. Future studies using BARseq to map the projections of neuronal types at cellular resolution, from multiple cortical areas, at multiple developmental time points, could help resolve the single-cell basis of the wire-by-similarity organization.

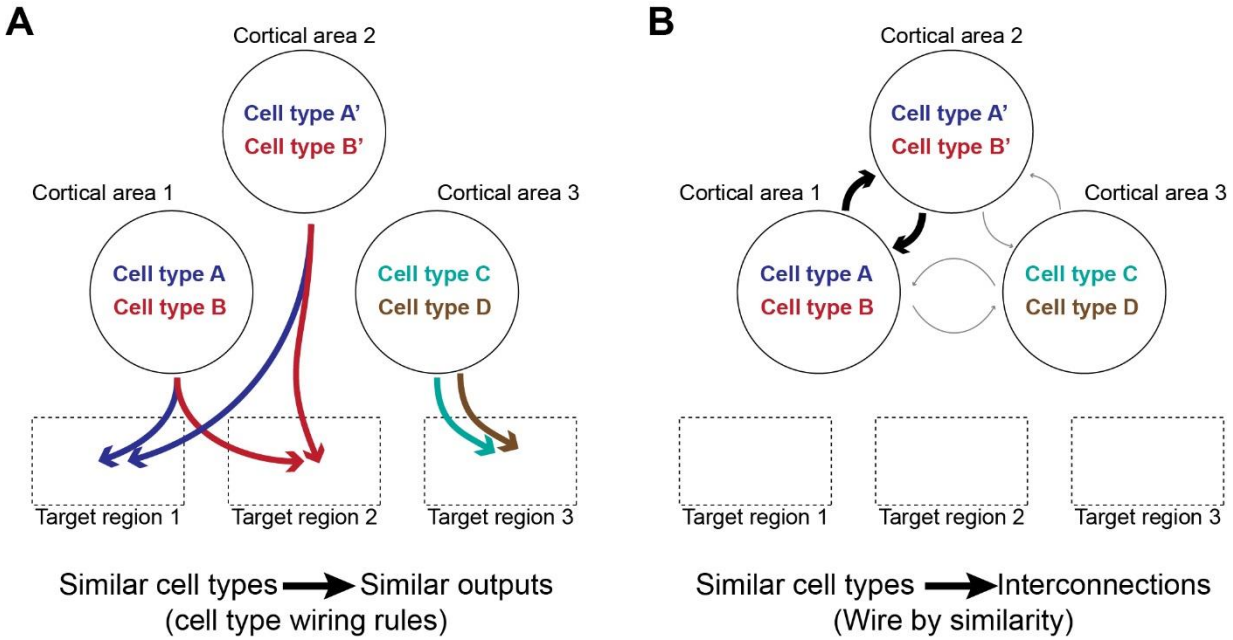


Figure 6. Comparison of wiring rules.

(A) *In cell type wiring rules, neurons of similar types usually have similar outputs. These wiring rules usually do not necessarily make specific predictions about interconnections among these neuronal types.* ***(B)*** *Across cortical areas, we find that areas with similar cell types are more interconnected. This model does not make predictions on the similarities of the outputs to other brain regions.*

Cortical modules may reflect distinct developmental processes

The modular organization of the cortex suggests that the cortex develops through at least two distinct processes, one that establishes distinct neuronal types on a broad spatial scale (i.e. across modules) and one that refines the composition of neuronal populations on a finer spatial scale (i.e. within modules). Strikingly, this view is reminiscent of two processes that are central to cortical development[14]: the "protomap," which is composed of intrinsic molecular programs, establishes the broad architecture of the cortex[58] independent of extrinsic signaling[61]; extrinsic mechanisms, including thalamocortical axons, then refine the borders between adjacent cortical areas[62]. If the modular organization of the cortex results from the interaction between these two developmental processes, then our results suggest that these two processes are distinct in their effects on the development of transcriptomic types: The protomap establishes the cortical modules and distinct sets of neuronal types, whereas the thalamocortical axons largely modify the composition of existing neuronal types within a module. Because BARseq is high throughput and low cost, it can be applied to multiple animals with different developmental perturbations to causally test this hypothesis on a whole-cortex scale. Establishing the developmental origin of cortical modules may reveal the biological origin of the wire-by-similarity organization and provide the missing link between the complex neuronal connectivity of the cortex and transcriptomic types.

Methods

Animals and tissue processing

All animal procedures were carried out in accordance with the Institutional Animal Care and Use Committee (protocol no. 19-16-10-07-03-00-4) at Cold Spring Harbor Laboratory. The animals were housed at maximum of five in a cage on a 12-h on/12-h off light cycle. The temperature in the facility was kept at 22 °C with a range not exceeding 20.5 °C to 26 °C. Humidity was maintained at around 45–55%, not exceeding a range of 30–70%. The animals used in the study were 7-8 week-old C57BL6/J animals purchased from Jackson Laboratory. To collect brains for BARseq, we euthanized the animals with isofluorane overdose, decapitated the animals, embedded the brain in OCT and snap-froze in an ethanol dry-ice bath. In experiments in which only one hemisphere was used, we bi-sected the brain along the midline first before OCT embedding. The brains were then sliced to 20 µm hemi-coronal sections and mounted onto Superfrost Plus Gold slides. Eight sections were mounted onto each slide.

BARseq library preparation

BARseq samples were prepared as previously described [23]. Briefly, slides were fixed in PFA, dehydrated in ethanol, rehydrated in PBST (PBS with 0.5% Tween-20), then reverse transcribed using random primers with a N-terminal amine group and Revert-aid H-minus reverse transcriptase (Thermo Fisher). On the second day, we crosslinked the cDNA products, then performed padlock probe hybridization and ligation, followed by rolling circle amplification. On the third day, we crosslinked rolonies. To sequence the samples, we hybridized sequencing primers and manually performed sequencing using Illumina HiSeq Rapid SBS kit v2 following previous protocols [23]. After seven sequencing cycles were completed, we then striped the sequencing primers using three incubations at 60℃ for 5 mins each in 40% formamide, 2× SSC, 0.1% Tween 20. We then hybridized fluorescent probes in 10% formamide, 2× SSC for 10 mins, followed by incubation in 2 µg/mL DAPI in PBST for 5 mins. We then imaged the final hybridization cycle. Detailed protocols are available at protocols.io: https://www.protocols.io/private/b6f33ccee34ba028cfe37ec9b868a9aa. Primers and padlock probes used are listed in Supplementary Table 1.

BARseq data collection

BARseq data were collected as described previously [23] on a Olympus IX-81 microscope with PI P-736 piezo z-stage, Olympus UCPLFLN20X 20× 0.7NA air objective, a Crest xlight v2 spinning disk confocal, 89North LDI-7 laser bank, and Photometrics BSI-prime camera. The filters used are listed in Supplementary Table 2. Because the microscope stage could only fit 3 slides at a time, the posterior 22 sections and the anterior 18 sections were collected in two separate batches.

BARseq data processing

BARseq data were processed as described previously[23] with slight modifications. Briefly, we created max-projections from the image stacks, applied noise reduction using Noise2Void [63], applied background subtraction, corrected for channel shift and bleedthrough, and registered images through all sequencing cycles. We segmented cells using Cellpose[43], decoded rolonies using BarDensr[40], and assigned rolonies to cells. Finally, the images were stitched to generate whole-slice images. All processing steps were performed

on each imaging field of view (FOV) separately, not on the stitched images, to avoid stitching artifacts, to avoid alignment artifacts due to imperfect optics, and to facilitate parallel processing. The stitched images were only used to generate a transformation, which we applied to each rolony and cell after all other steps were finished. Overlapping cells from neighboring FOVs were detected using a custom implementation of sort and sweep[64], an algorithm that is commonly used in detecting object collision in video games. When two or more cells were detected as overlapping, cells that had more read counts (which we assumed to have higher read quality) were kept, and all other cells were removed. See Data and Code Availability for processing script and intermediate processed data.

Quality control

After segmentation, we obtained a total of 2,265,631 segmented cells. We kept all cells expressing at least 5 unique genes and at least 20 total counts, resulting in a count matrix with 107 genes and 1,259,256 cells.

Iterative clustering and annotation of BARseq transcriptomic data

To obtain H1, H2, and H3 types, we adopted an iterative clustering pipeline adapted from single-cell RNA sequencing (scRNAseq) studies [65]. We performed 3 rounds of clustering with the following steps: normalization, dimension reduction (PCA), computation of a shared nearest neighbor (SNN) network, Louvain clustering. We normalized counts to CP10 (counts per 10) values, then applied the log1p transformation (log with a pseudo-count of 1). Because our panel contains markers, we skipped highly variable gene selection and ran PCA on all genes. We computed PCA using the scater::runPCA function and kept the first 30 PCs. We built the SNN network using the scran::buildSNNGraph function with 15 nearest neighbors and the "rank" metric. We ran the Louvain clustering as implemented in the igraph::cluster_louvain function with default parameters. All UMAP visualizations were obtained using the scater::runUMAP function starting from the PCA dimension reduction and using the 15 nearest neighbors.

In the first round of clustering, we separated cells in three classes reflecting neurotransmitter expression: excitatory (expressing *Slc17a7*), inhibitory (expressing *Gad1*), and others (expressing neither *Slc17a7* and *Gad1*). Because marker genes are frequently undetected at the single-cell level, we ran the clustering pipeline on all cells, obtaining 24 clusters, then assessed marker expression at the cluster level. From the UMAP visualization, we distinguished 3 groups of clusters. The first group contained excitatory clusters expressing *Slc17a7*, the second group contained inhibitory clusters expressing *Gad1*, the third group expressed neither of these markers. Based on these observations, we manually annotated the clusters as "excitatory", "inhibitory", and "other", respectively (H1 types, 642,340, 427,939, and 188,977 cells, respectively). In the second round of clustering, we extracted all cells labeled as "excitatory" and "inhibitory", then ran our pipeline again on each H1 type separately. By re-running the pipeline, the dimension reduction and clustering are better targeted at finding variability specific to either excitatory or inhibitory cells. We obtained 18 excitatory clusters and 11 inhibitory clusters, which formed the basis for our H2 types. In the third round of clustering, we re-ran the pipeline on each excitatory H2 type, obtaining roughly 5 to 6 clusters by type (117 total H3 types).

To annotate H2 and H3 types, we examined the brain-wide distribution and the marker expression of H2 and H3 types. Almost all H3 types showed brain-area specificity, suggesting that the data were clustered at a biologically meaningful granularity. We annotated isocortical H2 types based on aggregate marker expression (see Supplementary Note 1 for marker selection). We annotated non-isocortical H2 types based

on the localization of cells (e.g., hippocampal areas, thalamus, entorhinal cortex). When H2 types contained a mix of isocortical and other cells (e.g., "L6 IT-like"), we split the H2 type into multiple H2 types, one containing the isocortical cells (e.g., "L6 IT"), the others containing the other cells (e.g., "PIR L6 IT-like", "AON DL"). At the end, we obtained a list of 11 cortical H2 types and 23 non-cortical H2 types. After mapping to scRNAseq reference types, we noticed that four H3 types ("PT AUD", "PT P RSP/IT-like?", "RSP DL", "CT CTX A/L-V") were assigned to incorrect H2 types ("L5 IT", "RSP DL", "L5 IT", "L6b"); we manually corrected their H2 annotation (to "PT", "PT", "RSP DL", "CT").

Mapping of BARseq types to scRNAseq reference types

To map BARseq types to reference types, we used a k-nearest neighbor (kNN) approach to label each cell according to its closest neighbors in a whole-cortex and hippocampus reference compendium [13]. First, we evaluated the accuracy of the kNN approach on a subset of the reference compendium using leave-one-cell-out cross-validation (10X MOp dataset, "Glutamatergic" cells, cortical cells labeled as "CTX" or "Car3", clusters with ≥ 50 cells, CP10K and log1p normalization). To transfer labels, we picked each cell's closest 15 neighbors using the BiocNeighbors::queryKNN function, then predicted the cell's type by taking a majority vote across the neighbors. We compared accuracies across 4 gene panels: highly variable genes (HVGs, 2000 genes selected using scran::modelGeneVar and scran::getTopHVGs), HVG selection followed by PCA (30 components, scater::runPCA), the BARseq panel (107 genes, after excluding Gad1 and Slc17a7), the BARseq panel with reads downsampled to match BARseq sensitivity (107 genes, binomial sampling of reads, re-normalization through sample-wise ranking and scaling). For the latter gene set, reads were downsampled for each gene according to the sensitivity ratio (BARseq average counts divided by reference average counts). For genes that had a sensitivity ratio $r > 1$, we oversampled reads to match BARseq sensitivity (reads multiplied by $\lfloor r \rfloor$ + binomial sampling with probability $\lfloor r \rfloor - r$).

Having validated that the kNN mapping procedure was able to assign cell types with high accuracy, we applied the same procedure (sensitivity adjustment of reference datasets, sample-wise ranking and scaling of reference and target datasets, BiocNeighbors::queryKNN with 15 neighbors) to assign a reference label to each BARseq cell. In contrast with the previous evaluation, we used all excitatory cells from the reference compendium (40 individual datasets, all "Glutamatergic" cells) and adjusted reference reads using a simplified downsampling procedure (reads multiplied by the sensitivity ratio for each gene). To compute the overlap between BARseq and reference cell types, we used the Jaccard coefficient (number of cells labeled as BARseq type b and predicted to be reference type r divided by the number of cells of type b or predicted to be type r).

We mapped both H3 types in the cortex (Fig. 2E) and those in the hippocampal formation (ED Fig. 2F) to single-cell RNAseq data. However, we do not expect perfect matching for clusters outside of the cortex, because our dataset sampled additional brain regions that were not sampled in the single-cell RNAseq data. In addition, our gene panel was optimized for cortical excitatory neurons and could miss highly differentially expressed genes in other brain regions.

Variance of expression explained by H2 types, H3 types, and space

To evaluate how well H2 types, H3 types, and spatial information recapitulate the variability of expression, we performed one-way ANOVA on pseudo-bulk data. This analysis was run on a subset of data containing

the 8 isocortical H2 types with isocortex-wide localization ("L2/3 IT", "L4/5 IT", "L5 IT", "L6 IT", "PT", "NP", "CT", "L6b").

We started by computing the pseudo-bulk expression matrix B_{gts}, providing expression of gene g for H3 type t in spatial bin s. We defined 540 spatial bins containing an average of 14 cells per H3 type as follows: 27 slices along the A-P axis (corresponding to slices 7 to 33), 20 bins along the unwarped M-L axis for each slice (chosen to balance the number of cells in each bin, computed independently for each slice). Slices at the anterior and posterior end of the brain were excluded because coronal sections were not perpendicular to the cortical surface at these extreme positions and would thus bias gene expression. Starting from the gene by cell count matrix C_{gc}, we have $B_{gts} = mean_{c\in t, c\in s}\ (C_{gc})$.

Next, we computed the variance explained by the 8 H2 types, 51 H3 types and 540 spatial bins by applying the one-way ANOVA formula for each gene and factor independently. Let $M = mean_{t\in H3, s\in bin}\ (B_{gts})$ be the overall average expression and $T = \sum_{t\in H3, s\in bin}(B_{gts} - M)^2$ be the total variance. For gene g, we computed the variability explained as follows:

$$VE_{space} = \sum_{s\in bin} \quad (mean_{t\in H3}\ (B_{gts}) - M)^2\ /T$$

$$VE_{H3} = \sum_{s\in bin} \quad (mean_{s\in bin}\ (B_{gts}) - M)^2\ /T$$

$$VE_{H2} = \sum_{h\in H2} \quad (mean_{s\in bin, t\in h}\ (B_{gts}) - M)^2\ /T$$

Because H3 types are nested factors of H2 types, the variability explained by H3 types is necessarily higher; the additional variability explained by H3 types is given by $\Delta VE = VE_{H3} - VE_{H2}$.

Extraction of recurrent spatial patterns using non-negative matrix factorization

The ANOVA analysis revealed the presence of recurrent spatial patterning across genes and H2 types. We used non-negative matrix factorization to extract these patterns. First, we defined a pseudobulk matrix using the same procedure as the ANOVA analysis (see above), except that we computed the pseudobulk matrix at the level of H2 types. Starting from the count matrix C_{gc}, we have $B_{gts} = mean_{c\in t, c\in s}\ (C_{gc}\ )$, where t is one of the 8 H2 types. Here, we consider the spatial bins as features (rows), genes and types as variables (columns), resulting in a matrix with 540 rows and 848 columns. Because spatial patterns had different scales (average level of expression) across genes and H2 types, we rescaled each column using L2-normalization (squared columns sum to 1). This rescaling ensures that factors reflect recurrent patterns (seen in multiple genes and H2 types) rather than a single instance (highly expressing gene in one H2 type). This procedure (pseudobulking at H2 level and rescaling) can also be seen as a correction for H2 type composition (overall expression patterns are dominated by the most common or the highest expressing H2 type). We extracted 10 NMF factors using the NMF::nmf function using default parameters (Brunet algorithm), obtaining a nonnegative basis matrix W (540 bins by 10 factors) containing spatial patterns and a nonnegative coefficient matrix H (10 factors by 848 columns) such that B≈W.H. For later analysis, we removed 3 NMF factors that reflected obvious slice-specific batch effects (see Supplementary Note 4).

To identify genes associated with each NMF factor, we computed the average fraction of expression variance explained by each NMF factor. Given an NMF factor f, we computed the predicted gene expression for gene g in type t as $\widehat{B_{gt.}} = W_{.f}.H_{fgs}$, where $W_{.f}$ is the column in W representing factor f, and H_{fgs} is the coefficient associated with factor f, gene g and type t. The variance explained is then computed as $VE_t = (T - \sum_{s \in bin}(B_{gts} - \widehat{B_{gts}})^2)/T$, where $T = \sum_{s \in bin} {B_{gts}}^2$ is the total uncentered variance. We then took the average variance explained across the 8 H2 types. To estimate the null fraction of variance explained, we permuted coefficients associated with each factor (permutation of rows in H), then recomputed the average variance explained across all genes and factors. We labeled gene-NMF associations as "significant" if the average variance explained exceeded the 99th percentile of the null distribution.

To identify H3 types associated with each NMF factor, we computed two measures of association: enrichment of NMF-associated genes, correlation of spatial distribution with NMF patterns. Using the MetaMarkers package, we computed differentially expression (DE) statistics for each H3 type (1-vs-all differential expression against other H3 types from the same H2 type). We then asked whether top DE genes were enriched for NMF-associated genes (genes with significant association with a single NMF). We report the enrichment as an AUROC, asking how well expression fold changes predict NMF-associated genes. Independently, we computed the overlap between NMF patterns (columns in W) with the spatial distribution of H3 types (count of cells in bin s divided by total count) using the Spearman correlation. Because the dynamic range of the correlation coefficient depends on the sparsity of the pattern (sparser patterns tend to have a smaller range of correlation values), we report the association as the Z-scored Spearman correlation (Z-scoring across H3 types within a given H2 type and factor).

Predicting cortical areas and locations using gene expression and H3 type composition

We first binned cells in the cortex into cubelets, which were drawn separately on each coronal section, spanned all cortical layers, and were about 100 μm - 200 μm on the M-L axis along the curvature of the cortex. To draw the cubelets so that their medial and lateral borders were perpendicular to the layers, we manually drew matching points along the top and bottom surface of the cortex, especially at locations where the curvature of the cortex is extreme (e.g. the medial part of the cortex). This step thus separates both the bottom and top surfaces of the cortex into several segments. Within each segment, we then cut the two surfaces into the same number of smaller segments of roughly equal distance. Each cubelet was then defined by connecting the ends of the small segments. Slices at the anterior and posterior end of the brain were excluded because coronal sections were not perpendicular to the cortical surface at these extreme positions and would thus bias composition of neuronal populations. Unlike the spatial bins used in the NMF analysis, which had equal cell numbers within each slice but unequal widths, the cubelets had similar widths across all slices, but not necessarily similar cell numbers (ED Fig. 4A).

Because the cubelets are generated within each of the large segments, they may be slightly different in size across different segments. We thus normalized both H3 type counts and gene read counts within each cubelet by the total number of cells and/or gene reads for downstream analyses. For the A-P locations of each cubelet, we used the CCF coordinates directly. For the M-L locations, we calculated an unwarped coordinate along the cortex within each coronal section as follows. We connected the centroids of adjacent cubelets and defined the unwarped distance between two cubelets as the sum of all connected lines across all cubelets between the two. We then defined the zero position along this unwarped M-L axis as the point that was closest to the point in CCF where the midline of the brain intersects the top surface of the brain. Thus cubelets on the medial side have negative M-L coordinates, whereas those on the dorso-lateral side have positive M-L coordinates.

To predict the coordinates of each cubelet, we trained random forest regression models, each with 50 trees and an in-bag-fraction of 0.5. We used 100-fold cross-validation to evaluate the performance of the models. To evaluate the prediction performance using the composition of H3 types within each H2 type, we trained similar regression models using only the relevant H3 types and evaluated performance using 5-fold cross-validation. To predict cortical area labels in CCF, we first assign CCF area labels to each cubelet using its centroid location. We then trained random forest classifiers with 500 trees and an in-bag-fraction of 0.5, and evaluated the performance with 10-fold cross validation. All models were built in MATLAB using the TreeBagger function.

Correspondence between abrupt changes in the composition of H3 types and area borders defined by CCF

To find positions of abrupt changes in the composition of H3 types within each coronal section, we performed principal component analysis on the composition of H3 types and calculated the two-norm of the 1st derivative of the first five principal components along the unwarped M-L axis. We then convoluted the two-norm of the derivatives with a smoothing window of the shape [0.25, 0.5, 0.25]. We then looked for local peaks with prominence that is larger than half a standard deviation, and with the value at the peak higher than the median of the smoothed norm of the derivatives. Peaks are considered close to a CCF-defined border if it is within 150 μm from that border, calculated based on the CCF coordinates of the centroids of cubelets. All relevant analyses were performed in MATLAB.

Inferring cortical modules from H3 type composition

We only clustered areas with at least 7 cubelets. For each pair of areas, we built a support vector machine classifier with Lasso regularization to predict the area identity given H3 type composition. We calculated the area under the ROC curves following 5-fold cross validation. One minus the AUROC values were used to build a distance matrix among cortical areas. We then performed Louvain community detection using this distance matrix, which generated six clusters and three areas that did not cluster with any other areas. We then built a dendrogram of the six clusters and three areas using the medians of all pairwise distances between areas from each pair of clusters and ward linkage. We then manually cut the dendrogram to generate the five modules. The cortical flatmap of modules was drawn based on that in Harris et al., 2019[4], and color-coded manually based on this analysis.

To test how strongly the modularity of cortical areas is, we calculated the modularity of clusters with the following perturbations: (1) Randomly shuffle cortical area labels across all cubelets. (2) Randomly assign each cubelet with the area label within 1-cubelet distance with same probabilities. That is, for each cubelet, there is an equal chance of assigning the label of the cubelet itself, or the labels of the two cubelets adjacent to it. (3) Randomly assign each cubelet with the area label within 2-cubelet distance with the same probabilities. In addition, for the random shuffling control, we calculated modularity for both the clusters identified by Louvain clustering on the original data and the clusters identified by Louvain clustering on the shuffled data.

All relevant analyses were performed in MATLAB. The linear classifier was generated using the fitclinear function, and Louvain community detection was performed using a MATLAB implementation by Antoine Scherrer.

Data and code availability

Raw sequencing images are available from the Brain Image Library. Cell-level and rolony-level data and scripts used for both data processing and data analysis are provided at Mendeley data (doi; 10.17632/8bhhk7c5n9.1)(https://data.mendeley.com/datasets/8bhhk7c5n9/draft?a=f37296fe-dc00-46a6-9e69-46f4c3d0deec for preview) and on Github (https://github.com/gillislab/barseq_analysis).

Acknowledgement

The authors thank W. Wadolowski for technical support; B. Tasic, Z. Yao, B. Long, D-W. Kim, M. Rue, and other members of the Allen Institute for discussion. This work was supported by the National Institutes of Health (5RO1NS073129, 5RO1DA036913, RF1MH114132 and U01MH109113 to A.M.Z; R01MH113005 and R01LM012736 to J.G.; and U19MH114821 to A.M.Z. and J.G.; 1DP2MH132940 to X.C.), the Brain Research Foundation (BRF-SIA-2014-03 to A.M.Z.), IARPA MICrONS (D16PC0008 to A.M.Z.), and Robert Lourie award (to A.M.Z.). A.M.Z. was supported by an Allen Distinguished Investigator Award, a Paul G. Allen Frontiers Group advised grant of the Paul G. Allen Family Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. X.C. and A.Z. wish to thank the Allen Institute founder, Paul G. Allen, for his vision, encouragement, and support.

Competing interests

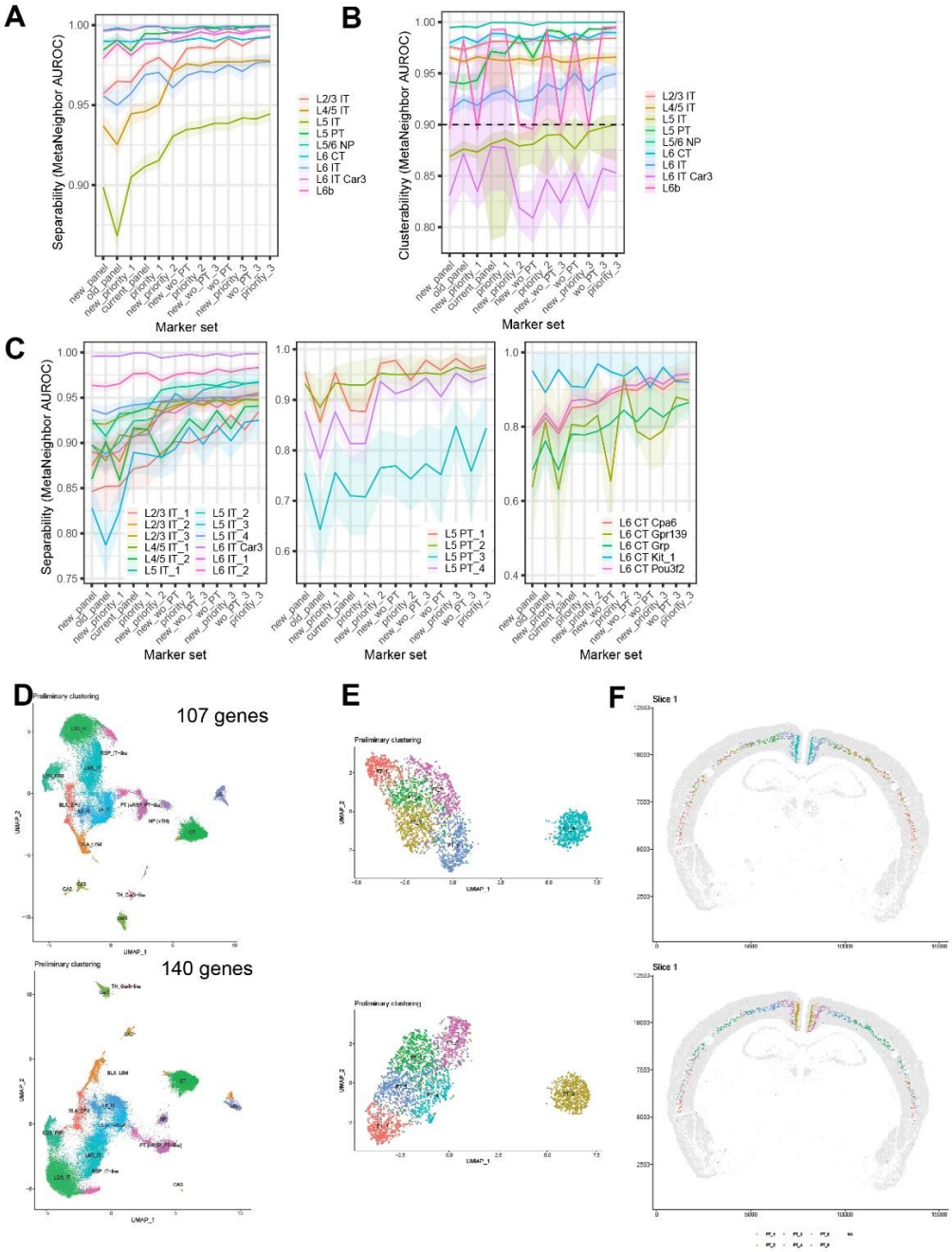
A.M.Z. is a founder and equity owner of Cajal Neuroscience and a member of its scientific advisory board. The remaining authors declare no competing interests**.**

Author contribution

X.C. conceived the study, S.F. selected the gene panel, X.C. collected data, X.C., S.F., J.G., A.M.Z., and A.Z. analyzed the data, X.C., S.F., J.G., and A.M.Z. wrote the paper.

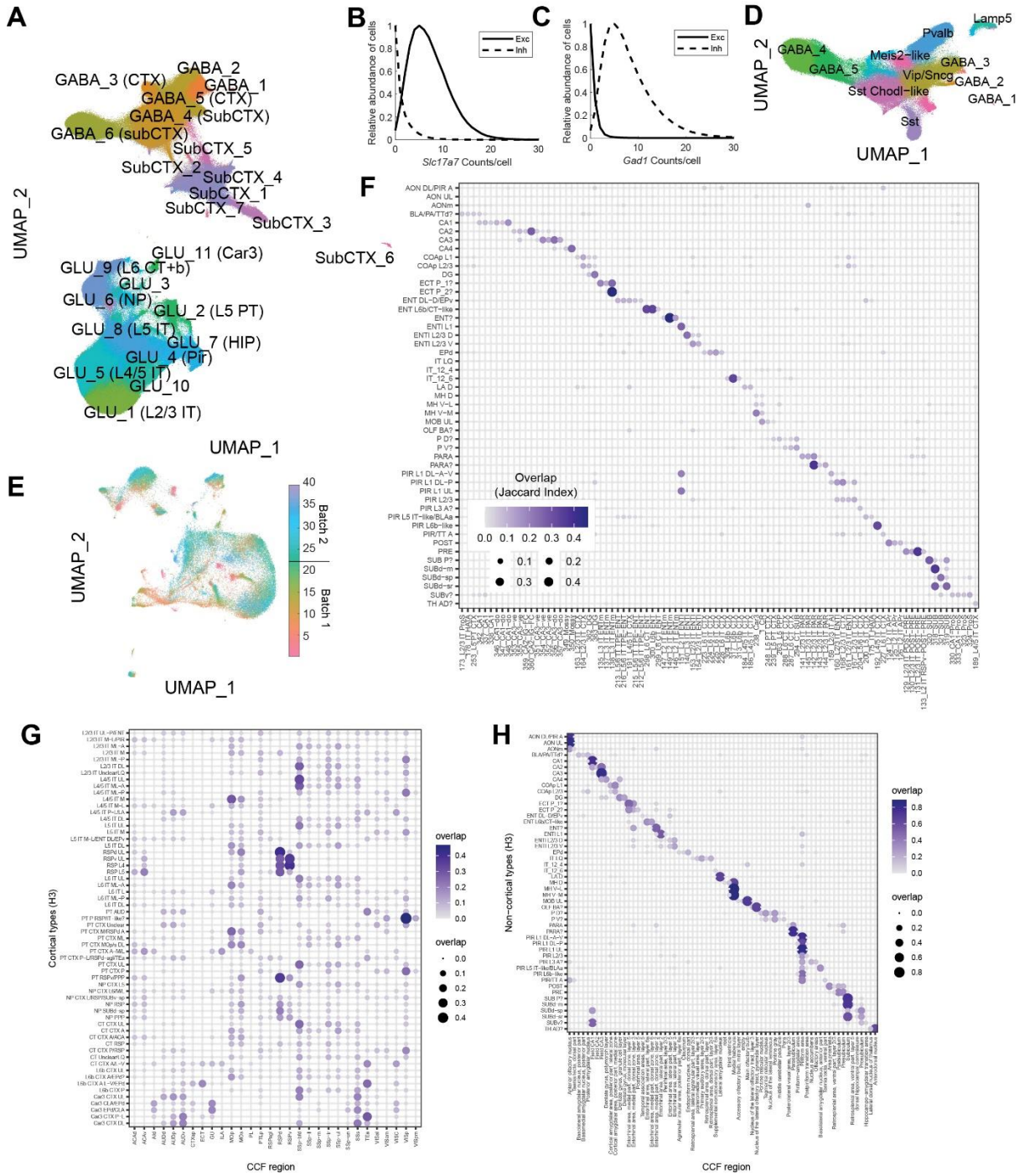
Materials and correspondence

Correspondence and requests for materials should be addressed to X.C.



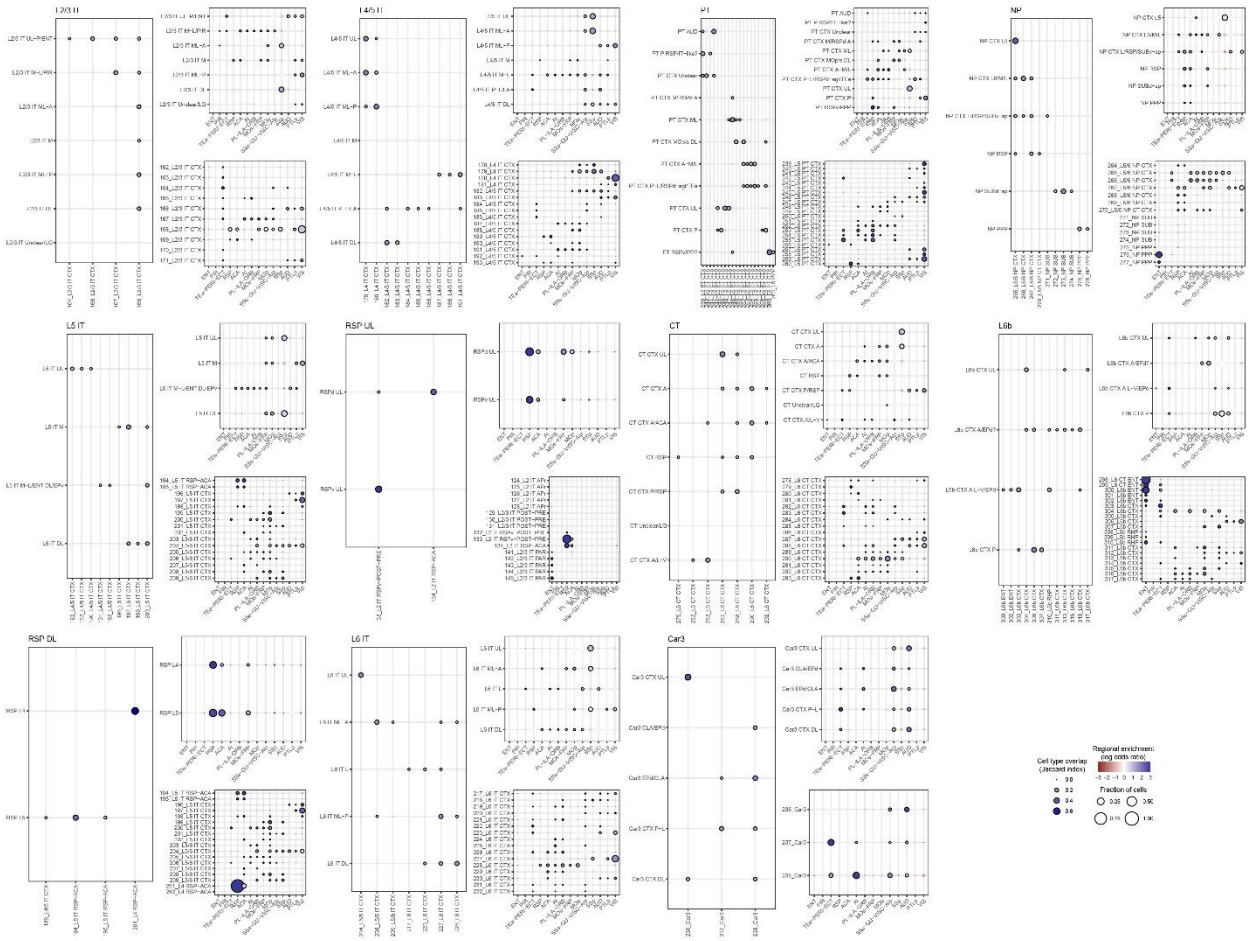
ED Figure 1. In silico design and evaluation of the gene panel.

(A) In silico assessment of H2 type separability (supervised analysis, ability to distinguish cell types given reference labels) in a surrogate snRNAseq dataset across 13 potential marker sets ranging from 45 to 107 genes. **(B)** In silico assessment of H2 type clusterability (unsupervised analysis, ability to recover reference labels through standard clustering) across the marker sets. **(C)** In silico assessment of H3 type separability for IT, L5 ET (PT), and L6 CT cells across the marker sets. **(D)(E)** UMAP plots of gene expression of cortical excitatory neurons (D) and L5 ET neurons (E) calculated from the 107-gene panel with or without an additional 33 genes. Colors indicate H2 types in (D) and H3 types in (E). **(F)** The distribution of L5 ET types within a coronal section.



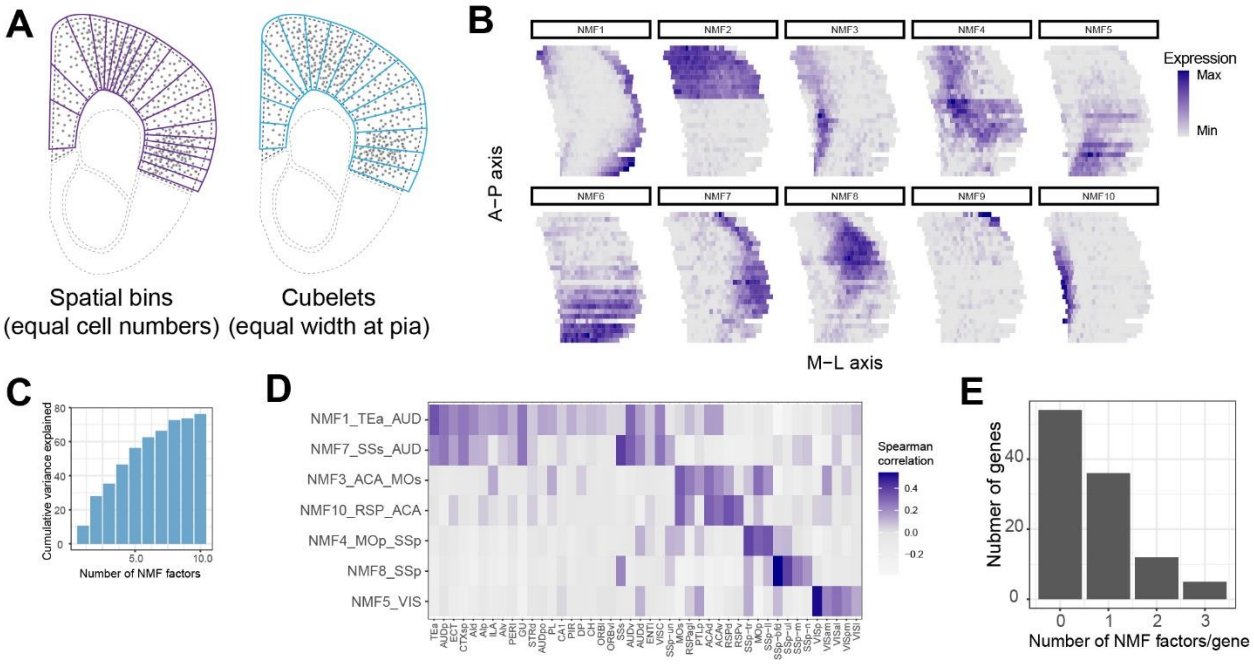
ED Figure 2. Hierarchical clustering of BARseq data.

(**A**) UMAP plot of the gene expression of all cells. Colors and labels indicate H1 clusters. (**B**)(**C**) Histograms of *Slc17a7* and *Gad1* counts per cell in excitatory and inhibitory neurons. (**D**) UMAP plot of the gene expression of inhibitory neurons. Colors and labels indicate H2 types of inhibitory neurons. (**E**) UMAP plot of the gene expression of excitatory neurons. Colors indicate slice numbers. The coordinates of dots in the UMAP plot are the same as those in Fig. 2C. (**F**) Cluster correspondence between non-isocortical H3 types in BARseq (rows) and single-cell RNAseq (columns)[13]. (**G**)(**H**) Overlap between isocortical (G) and non-isocortical (H) H3 types and CCF-defined cortical areas.



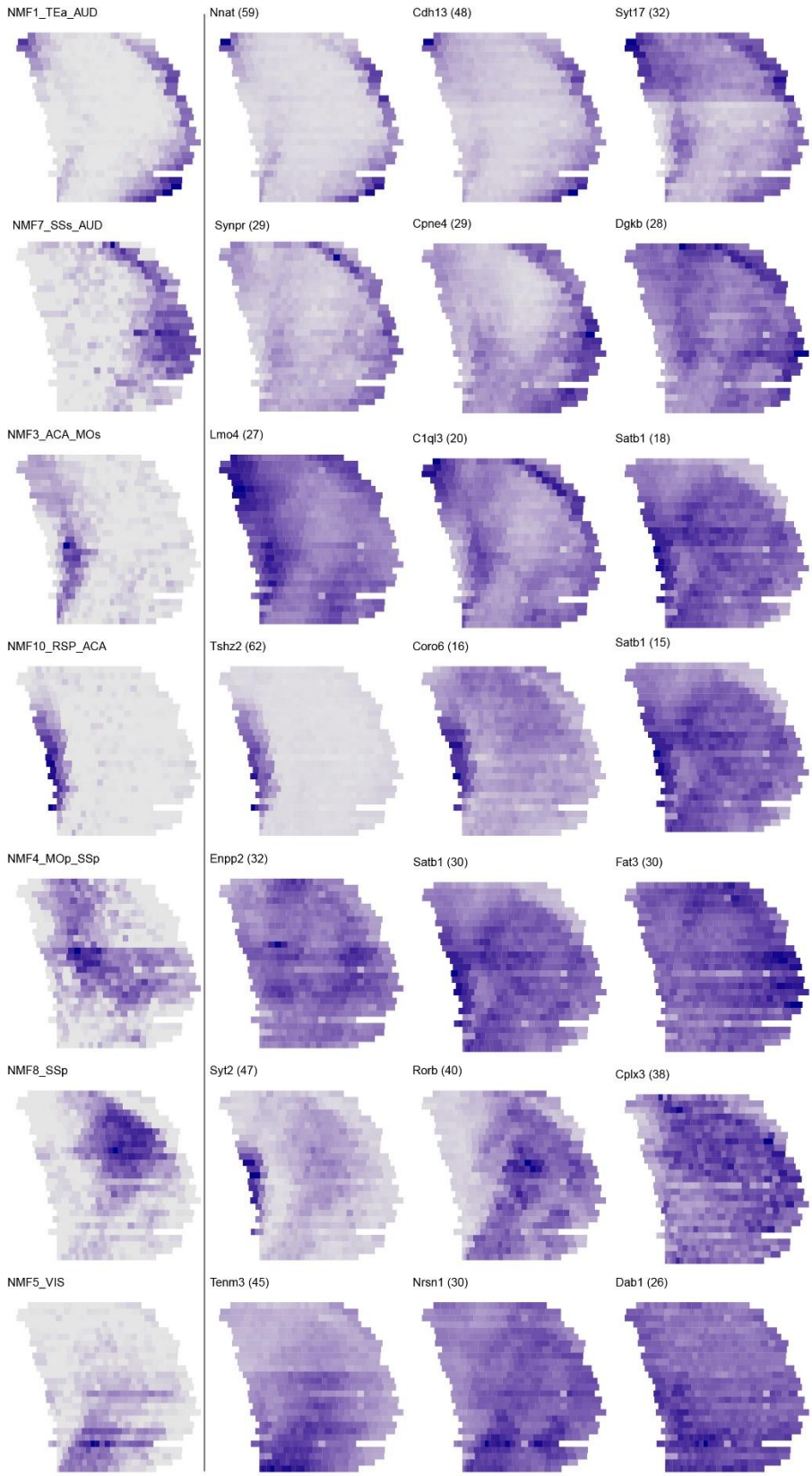
ED Figure 3. Mapping and comparative regional enrichment of BARseq and scRNAseq types.

For each BARseq H2 type, we show the mapping of BARseq H3 types with reference scRNAseq type (*left*), the CCF enrichment of H3 types (*top right*), and the CCF enrichment of scRNAseq types (*bottom left*). The mapping between BARseq and scRNAseq types is quantified as the Jaccard Index, significant associations (permutation test) are shown by outlining dots with black circles. The regional enrichment is quantified as an odds ratio, significant deviations (hypergeometric test) are shown by outlining dots with black circles. In the mapping panel, all BARseq H3 types are shown but, for readability, only scRNAseq types with significant associations are plotted. In contrast, the CCF enrichment is shown for all scRNAseq types belonging to subclasses that are equivalent to the BARseq H2 type (e.g., the BARseq L4/5 IT type corresponds to the L4 IT and L4/5 IT subclasses in the scRNAseq dataset).



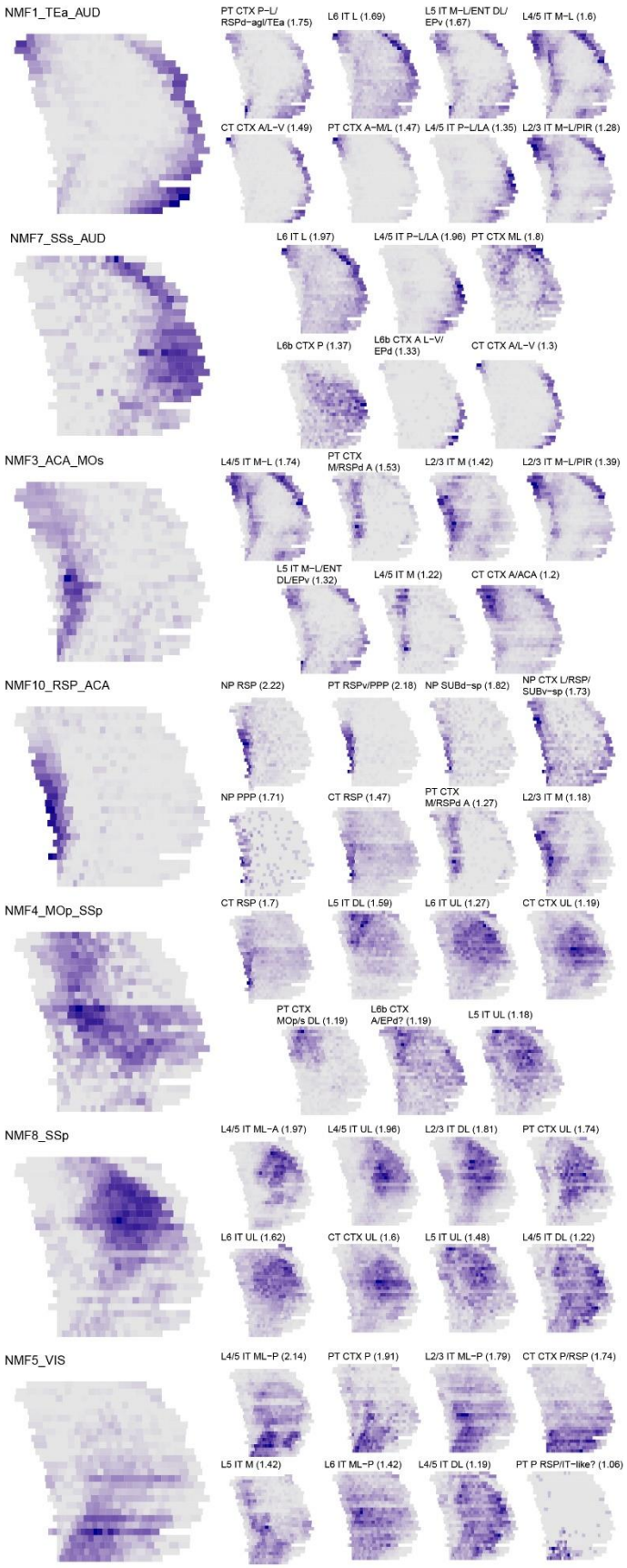
ED Figure 4. Identifying shared spatial patterns of gene expression using non-negative matrix factorization

(A) Illustrations of the definition of "spatial bins" used in gene expression analyses (purple outlines) and "cubelets" used in the analyses of H3 type distributions (blue outlines). The definition of Spatial bins aimed for equal cell numbers across bins within a slice, whereas the definition of cubelets aimed for equal width on the surface of the cortex. Dots indicate cells. (**B**) Spatial patterns of all 10 NMF factors. (**C**) Cumulative variance explained by the indicated number of NMF factors. (**D**) Spearman correlation between NMF factors and cortical areas. (**E**) Histogram of the number of NMF factors that each gene is associated with.



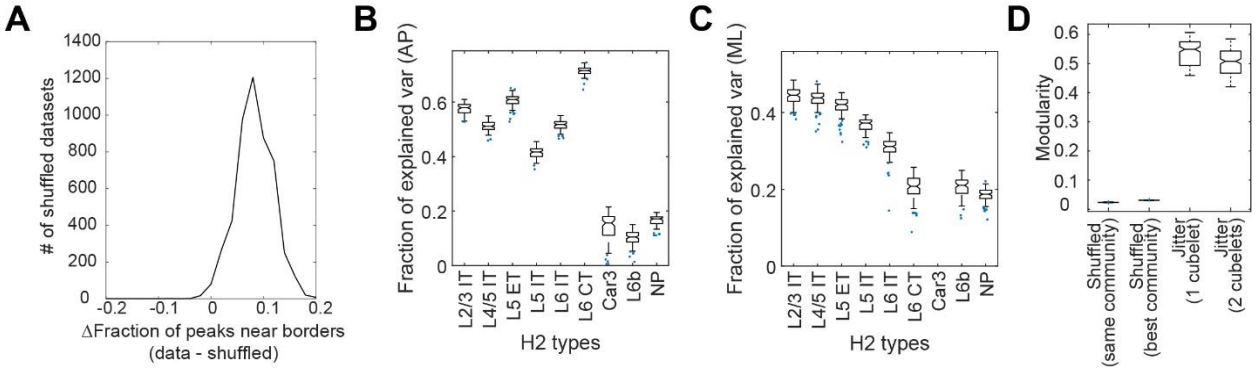
ED Figure 5. Expression patterns of top genes associated with each NMF spatial pattern.

The left column shows the pattern associated with each NMF; the right column shows the overall expression patterns (total expression counts across all cortical cells) of the top 3 genes associated with each NMF. Expression patterns were min-max standardized (max expression = blue). Numbers in parentheses next to gene names show the average percentage of gene expression variance explained by the NMF pattern across cortical H2 types.



ED Figure 6. Distribution patterns of top H3 types associated with each NMF spatial pattern.

The left column shows the pattern associated with each NMF; the right column shows the distribution pattern (fraction of cells from the H3 type found in each bin) of H3 types showing above-null association with the NMF pattern. Numbers in parenthesis next to H3 type names show the scaled Spearman correlation between the NMF pattern and the distribution pattern.



ED Figure 7. Cortical areas are distinct in H3 type composition.

(**A**) The positions of abrupt changes in the composition of H3 types were shuffled randomly within each slice, and the difference in the fractions of positions that were close to a CCF area border between the real data and shuffled data was calculated (see Methods). Positive values indicate that abrupt changes in the composition of H3 types were more likely to be associated with area borders in real data than in shuffled control. This shuffling was repeated 5,000 times, and the distribution of this difference is plotted in a histogram. (**B**)(**C**) The composition of H3 types within each indicated H2 types (x-axes) were used to predict the AP (B) and ML (C) locations of a cubelet. For each H2 type, we performed 100 trials. In each trial, we randomly held 10% of data as test set to determine the fractions of variance explained. (**D**) The distribution of modularity of shuffled data, or data with 1-2 cubelets of jitter in CCF registration. For shuffled data, we calculated modularity based on either the same clusters obtained from real data, or by the best clusters obtained by Louvain community detection on the shuffled data.

References

1 Von Bonin, G. The neocortex of Macaca mulatta. *Monographs in Medical Sciences*. **5**, 136 (1947).

2 Vogt, C. & Vogt, O. *Allgemeine ergebnisse unserer hirnforschung*. Vol. 25 (JA Barth, 1919).

3 Brodmann, K. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. (Barth, 1909).

4 Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195-202, doi:10.1038/s41586-019-1716-z (2019).

5 Zingg, B. *et al.* Neural networks of the mouse neocortex. *Cell* **156**, 1096-1111, doi:10.1016/j.cell.2014.02.023 (2014).

6 Bertolero, M. A., Yeo, B. T. & D'Esposito, M. The modular and integrative functional architecture of the human brain. *Proc Natl Acad Sci U S A* **112**, E6798-6807, doi:10.1073/pnas.1510619112 (2015).

7 Schwarz, A. J., Gozzi, A. & Bifone, A. Community structure and modularity in networks of correlated brain activity. *Magn Reson Imaging* **26**, 914-920, doi:10.1016/j.mri.2008.01.048 (2008).

8 Ferrarini, L. *et al.* Hierarchical functional modularity in the resting-state human brain. *Hum Brain Mapp* **30**, 2220-2231, doi:10.1002/hbm.20663 (2009).

9 He, Y. *et al.* Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS One* **4**, e5226, doi:10.1371/journal.pone.0005226 (2009).

10 Meunier, D., Lambiotte, R. & Bullmore, E. T. Modular and hierarchically modular organization of brain networks. *Front Neurosci* **4**, 200, doi:10.3389/fnins.2010.00200 (2010).

11 Huang, L. *et al.* BRICseq Bridges Brain-wide Interregional Connectivity to Neural Activity and Gene Expression in Single Animals. *Cell*, doi:10.1016/j.cell.2020.05.029 (2020).

12 Yao, Z. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598**, 103-110, doi:10.1038/s41586-021-03500-8 (2021).

13 Yao, Z. *et al.* A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* **184**, 3222-3241 e3226, doi:10.1016/j.cell.2021.04.021 (2021).

14 Cadwell, C. R., Bhaduri, A., Mostajo-Radji, M. A., Keefe, M. G. & Nowakowski, T. J. Development and Arealization of the Cerebral Cortex. *Neuron* **103**, 980-1004, doi:10.1016/j.neuron.2019.07.009 (2019).

15 Hobert, O. Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. *Proc Natl Acad Sci U S A* **105**, 20067-20071, doi:10.1073/pnas.0806070105 (2008).

16 Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat Rev Neurosci* **18**, 530-546, doi:10.1038/nrn.2017.85 (2017).

17 Stam, C. J., Jones, B. F., Nolte, G., Breakspear, M. & Scheltens, P. Small-world networks and functional connectivity in Alzheimer's disease. *Cereb Cortex* **17**, 92-99, doi:10.1093/cercor/bhj127 (2007).

18 Yao, Z. *et al.* Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease. *PLoS Comput Biol* **6**, e1001006, doi:10.1371/journal.pcbi.1001006 (2010).

19 Lynall, M. E. *et al.* Functional connectivity and brain networks in schizophrenia. *J Neurosci* **30**, 9477-9487, doi:10.1523/JNEUROSCI.0333-10.2010 (2010).

20 Fingelkurts, A. A. *et al.* Impaired functional connectivity at EEG alpha and theta frequency bands in major depression. *Hum Brain Mapp* **28**, 247-261, doi:10.1002/hbm.20275 (2007).

21 Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72-78, doi:10.1038/s41586-018-0654-5 (2018).

22 Chen, Y. *et al.* Wiring logic of the early rodent olfactory system revealed by high-throughput sequencing of single neuron projections. *bioRxiv*, 2021.2005.2012.443929, doi:10.1101/2021.05.12.443929 (2021).

23 Sun, Y. C. *et al.* Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections. *Nat Neurosci* **24**, 873-885, doi:10.1038/s41593-021-00842-4 (2021).

24 Ke, R. *et al.* In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10**, 857-860, doi:10.1038/nmeth.2563 (2013).
25 Qian, X. *et al.* Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat Methods* **17**, 101-106, doi:10.1038/s41592-019-0631-4 (2020).
26 Bugeon, S. *et al.* A transcriptomic axis predicts state modulation of cortical interneurons. *Nature* **607**, 330-338, doi:10.1038/s41586-022-04915-7 (2022).
27 Chen, X. *et al.* High-throughput mapping of long-range neuronal projection using in situ sequencing. *Cell* **179**, 772-786, doi:10.1016/j.cell.2019.09.023 (2019).
28 Munoz-Castaneda, R. *et al.* Cellular anatomy of the mouse primary motor cortex. *Nature* **598**, 159-166, doi:10.1038/s41586-021-03970-w (2021).
29 Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090, doi:10.1126/science.aaa6090 (2015).
30 Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932-935, doi:10.1038/s41592-018-0175-z (2018).
31 Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342-357, doi:10.1016/j.neuron.2016.10.001 (2016).
32 Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* **39**, 313-319, doi:10.1038/s41587-020-0739-1 (2021).
33 Stahl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82, doi:10.1126/science.aaf2403 (2016).
34 Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792 e1721, doi:10.1016/j.cell.2022.04.003 (2022).
35 Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138-1142, doi:10.1126/science.aaa1934 (2015).
36 Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014 e1022, doi:10.1016/j.cell.2018.06.021 (2018).
37 Paul, A. *et al.* Transcriptional Architecture of Synaptic Communication Delineates GABAergic Neuron Identity. *Cell* **171**, 522-539 e520, doi:10.1016/j.cell.2017.08.032 (2017).
38 O'Leary, D. D., Chou, S. J. & Sahara, S. Area patterning of the mammalian cortex. *Neuron* **56**, 252-269, doi:10.1016/j.neuron.2007.10.010 (2007).
39 Lein, E. S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176, doi:10.1038/nature05453 (2007).
40 Chen, S. *et al.* BARcode DEmixing through Non-negative Spatial Regression (BarDensr). *PLoS Comput Biol* **17**, e1008256, doi:10.1371/journal.pcbi.1008256 (2021).
41 Wang, Q. *et al.* The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181**, 936-953 e920, doi:10.1016/j.cell.2020.04.007 (2020).
42 Puchades, M. A., Csucs, G., Ledergerber, D., Leergaard, T. B. & Bjaalie, J. G. Spatial registration of serial microscopic brain images to three-dimensional reference atlases with the QuickNII tool. *PLoS One* **14**, e0216796, doi:10.1371/journal.pone.0216796 (2019).
43 Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *bioRxiv*, 2020.2002.2002.931238, doi:10.1101/2020.02.02.931238 (2020).
44 Cadwell, C. R. *et al.* Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat Biotechnol* **34**, 199-203, doi:10.1038/nbt.3445 (2016).
45 Scala, F. *et al.* Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature* **598**, 144-150, doi:10.1038/s41586-020-2907-3 (2021).
46 Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111-119, doi:10.1038/s41586-021-03465-8 (2021).
47 Liu, J. *et al.* Concordance of MERFISH Spatial Transcriptomics with Bulk and Single-cell RNA Sequencing. *bioRxiv*, 2022.2003.2004.483068, doi:10.1101/2022.03.04.483068 (2022).

48 Zhang, M. *et al.* Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137-143, doi:10.1038/s41586-021-03705-x (2021).

49 Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun* **9**, 884, doi:10.1038/s41467-018-03282-0 (2018).

50 Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030 e1016, doi:10.1016/j.cell.2018.07.028 (2018).

51 Meyer, H. S. *et al.* Inhibitory interneurons in a cortical column form hot zones of inhibition in layers 2 and 5A. *Proc Natl Acad Sci U S A* **108**, 16807-16812, doi:10.1073/pnas.1113648108 (2011).

52 Sahara, S., Yanagawa, Y., O'Leary, D. D. & Stevens, C. F. The fraction of cortical GABAergic neurons is constant from near the start of cortical neurogenesis to adulthood. *J Neurosci* **32**, 4755-4761, doi:10.1523/JNEUROSCI.6412-11.2012 (2012).

53 Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* **19**, 335-346, doi:10.1038/nn.4216 (2016).

54 Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, doi:10.1126/science.aat5691 (2018).

55 Lu, S. *et al.* Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain. *PLoS Biol* **19**, e3001341, doi:10.1371/journal.pbio.3001341 (2021).

56 Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791, doi:10.1038/44565 (1999).

57 Custo Greig, L. F., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat Rev Neurosci* **14**, 755-769, doi:10.1038/nrn3586 (2013).

58 Rakic, P. Specification of cerebral cortical areas. *Science* **241**, 170-176, doi:10.1126/science.3291116 (1988).

59 Campagnola, L. *et al.* Local connectivity and synaptic dynamics in mouse and human neocortex. *Science* **375**, eabj5861, doi:10.1126/science.abj5861 (2022).

60 Klingler, E. *et al.* Temporal controls over inter-areal cortical projection neuron fate diversity. *Nature* **599**, 453-457, doi:10.1038/s41586-021-04048-3 (2021).

61 Miyashita-Lin, E. M., Hevner, R., Wassarman, K. M., Martinez, S. & Rubenstein, J. L. Early neocortical regionalization in the absence of thalamic innervation. *Science* **285**, 906-909, doi:10.1126/science.285.5429.906 (1999).

62 Chou, S. J. *et al.* Geniculocortical input drives genetic distinctions between primary and higher-order visual areas. *Science* **340**, 1239-1242, doi:10.1126/science.1232806 (2013).

63 Krull, A., Buchholz, T.-O. & Jug, F. Noise2Void - Learning Denoising from Single Noisy Images. arXiv:1811.10980 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv181110980K>.

64 Baraff, D. & Witkin, A. Dynamic simulation of non-penetrating flexible bodies. *SIGGRAPH Comput. Graph.* **26**, 303–308, doi:10.1145/142920.134084 (1992).

65 Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* **17**, 137-145, doi:10.1038/s41592-019-0654-x (2020).