nature
biotechnology

Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence *in situ* hybridization data

Qian Zhu[1], Sheel Shah[2,3], Ruben Dries[1], Long Cai[2] & Guo-Cheng Yuan[1]

How intrinsic gene-regulatory networks interact with a cell's spatial environment to define its identity remains poorly understood. We developed an approach to distinguish between intrinsic and extrinsic effects on global gene expression by integrating analysis of sequencing-based and imaging-based single-cell transcriptomic profiles, using cross-platform cell type mapping combined with a hidden Markov random field model. We applied this approach to dissect the cell-type- and spatial-domain-associated heterogeneity in the mouse visual cortex region. Our analysis identified distinct spatially associated, cell-type-independent signatures in the glutamatergic and astrocyte cell compartments. Using these signatures to analyze single-cell RNA sequencing data, we identified previously unknown spatially associated subpopulations, which were validated by comparison with anatomical structures and Allen Brain Atlas images.

Human and other multicellular organisms are composed of diverse cell types that are characterized by distinct gene expression patterns. In each cell type, there is also considerable heterogeneity. The source of cellular heterogeneity remains poorly understood, but it is commonly thought to be modulated by the balance between intrinsic regulatory networks and extrinsic cellular microenvironment[1–5]. Recently, the rapid development of single-cell technologies has enabled accurate and simultaneous measurements of cell position and gene expression[6–9], thereby providing an opportunity to systematically characterize cellular heterogeneity. However, the relative contributions of intrinsic and extrinsic factors in mediating cell-state variation remain poorly understood.

Currently, there are two major, complementary approaches for single-cell transcriptomic profiling. The first is single-cell RNA sequencing (scRNAseq)[6,8,10–15]. By combining single-cell isolation, library amplification and massively parallel sequencing, scRNAseq provides the most comprehensive view of transcriptomes. The second approach is single-molecule fluorescence *in situ* hybridization (smFISH)[7,16–20], which can be used to detect mRNA transcripts with high sensitivity while maintaining the spatial information. Each technology features a distinct set of advantages and limitations. The sequential smFISH technology has the advantage of measuring the transcriptome with high accuracy in its native spatial environment, but current implementations profile only a few hundred genes, whereas scRNAseq provides whole-transcriptome estimation, but requires cells to be removed from their spatial environment, resulting in a loss of spatial information[19,21].

To combine the benefits of both technologies, we developed a computational approach to integrate scRNAseq and sequential smFISH. First, we used the scRNAseq data as a guide to accurately determine the cell types corresponding to the cells profiled by sequential smFISH. Second, we systematically detected distinct spatial domain patterns from sequential smFISH data. These spatial patterns were then in turn used to dissect the environment-associated variation in a scRNAseq data set.

This integrated approach allowed us to systematically dissect the respective contribution of spatially and cell-type-dependent factors in mediating cell-state variation (**Fig. 1a**), which has eluded previous studies. We analyzed the mouse visual cortex region and found that cell type differences represent only one component in cell-state variation, whereas the spatial environment had a substantial role in mediating gene activities, probably through cell-cell interactions (**Fig. 1a**) and signaling. Our integrated approach will be broadly applicable to the analysis of diverse tissues from various model systems.

RESULTS

Mapping scRNAseq cell types on seqFISH data

Given that scRNAseq, as a whole transcriptomic approach, can provide signatures for a diverse set of cell types, we took advantage of the whole-transcriptomic information obtained from scRNAseq data and developed a supervised cell type mapping approach by integrating seqFISH and scRNAseq data (**Fig. 1b**). Our goal differed from that of previous studies[22–26], where scRNAseq data were mapped onto conventional ISH images to predict cell locations. Traditional ISH images are not multiplexed or single-cell resolution. In a seqFISH experiment, transcripts from hundreds of genes are detected directly in individual cells in their native spatial environment at single-molecule resolution.

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [2]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA. [3]UCLA-Caltech Medical Scientist Training Program, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, California, USA. Correspondence should be addressed to L.C. (lcai@caltech.edu) or G.-C.Y. (gcyuan@jimmy.harvard.edu).

Received 21 December 2017; accepted 27 August 2018; published online 29 October 2018; doi:10.1038/nbt.4260

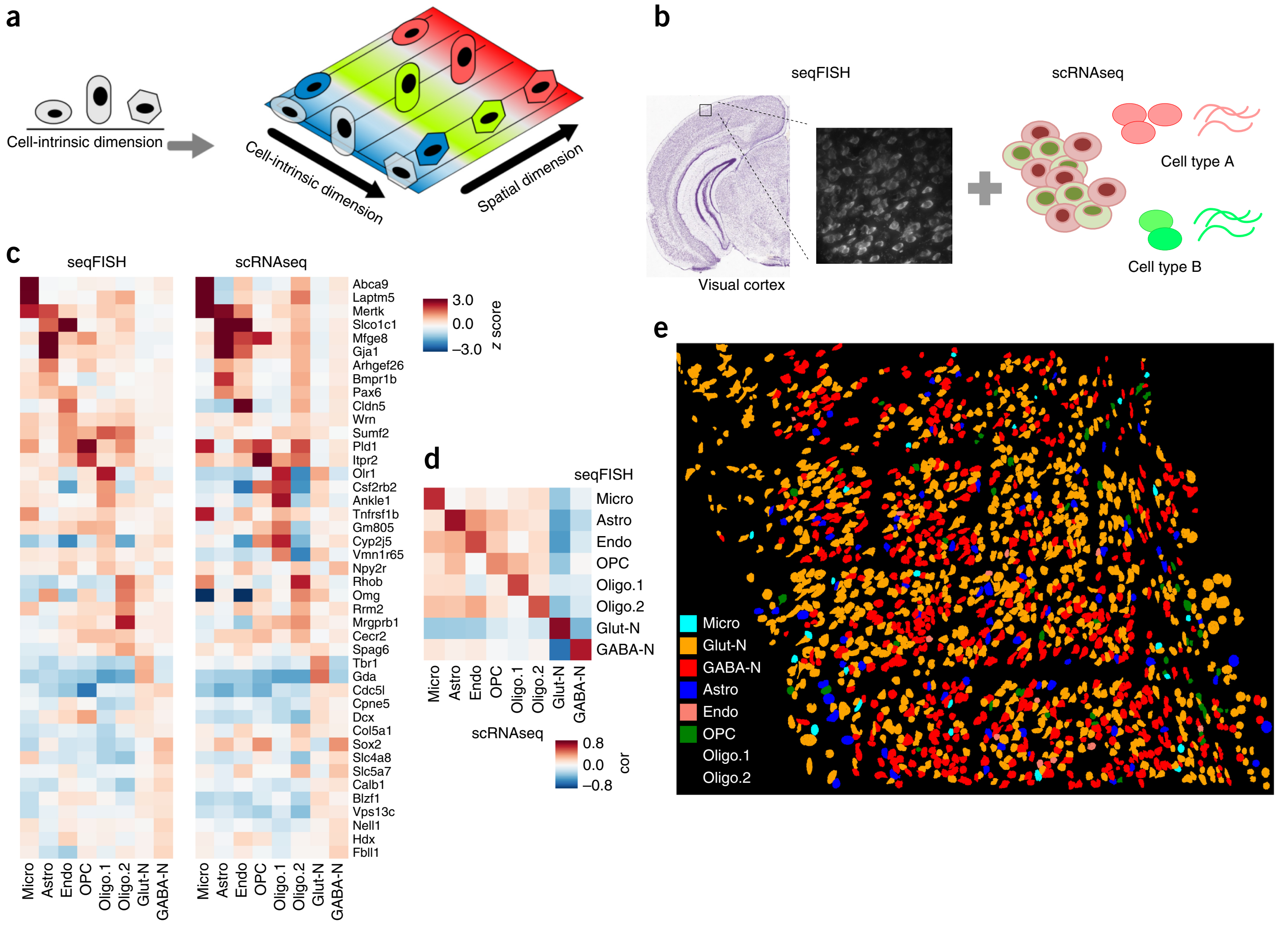

Figure 1 Overall goal of the project and cell type prediction in seqFISH data. (**a**) Cellular heterogeneity is driven by both cell type (indicated by shape) and environmental factors (indicated by colors). scRNAseq-based studies can only detect cell-type-related variation, as spatial information is lost. (**b**) Our goal was to decompose the contributions of each factor by developing methods to integrate scRNAseq and seqFISH data. (**c**) Prediction results evaluated by the comparison of cell type average expression profile across technologies for eight major cell types. Values represent expression *z* scores. SVM was tuned for the parameter C, which was set to 1 × 10^{-5} to optimize the cross-platform cell type to cell type correlations. The major cell types in the scRNAseq data set (Astro, *n* = 43 cells; Endo, *n* = 29; GABA-N, *n* = 761; Glut-N, *n* = 812; Micro, *n* = 22; OPC, *n* = 19; Oligo.1, *n* = 6; Oligo.2, *n* = 31) mapped to 97, 11, 556, 859, 22, 8, 21 and 23 cells in the seqFISH data set. (**d**) Pearson correlation between reference and predicted cell type averages ranged from 0.75 to 0.95. (**e**) Integration of seqFISH and scRNAseq data (illustrated in **b**) enabled cell type mapping with spatial information in the adult mouse visual cortex. Each cell type is labeled by a different color. Cell shape information was obtained from segmentation of cells from images (Online Methods). One mouse brain was assayed by seqFISH because of experimental cost.

Our strategy was to use scRNAseq data to capture the large cell type differences and then further investigate spatial patterning beyond cell type variations. We analyzed a published scRNAseq data set targeting the mouse visual cortex regions[27]. Eight major cell types, GABAergic, glutamatergic, astrocytes, three oligodendrocyte groups, microglia and endothelial cells, were identified from scRNAseq analysis[27]. To estimate the minimal number of genes that are required for accurate cell type mapping, we randomly selected a subset from the list of differentially expressed genes across these cell types and applied a multiclass support vector machine (SVM)[28,29] model using only the expression levels of these genes. The performance was evaluated by cross-validation. By using only 40 genes, we were able to achieve an average level of 89% mapping accuracy. Increasing the number of genes led to better performance (92% for 60 genes and 96% for 80 genes). Thus, there is substantial redundancy in transcriptomic profiles, which can be compressed into fewer than 100 genes.

We then investigated a seqFISH data set for the mouse visual cortex area[19]. We imaged a 1-mm × 1-mm contiguous area of the mouse visual cortex with four barcoded rounds of hybridization to decode 100 unique transcripts, followed by five rounds of non-combinatorial hybridization to quantify 25 highly expressed genes (**Supplementary Table 1**). These rounds of imaging were preceded by imaging of the DAPI stain in the region and followed by imaging of the Nissl stain to stain neurons in the region. The images were aligned and transcripts decoded as described previously[19]. Transcripts were assigned to cells that were segmented on the basis of Nissl and DAPI staining. We were able to quantify the expression levels of these 125 genes with high accuracy in a total of 1,597 cells.

After identifying differentially expressed genes across the 8 major cell types in a previous study[27], we selected the top 43 ($P < 10^{-20}$) of these 125 genes for cell type classification. These genes contained both highly expressed (>50 copies per cell) and lowly expressed genes

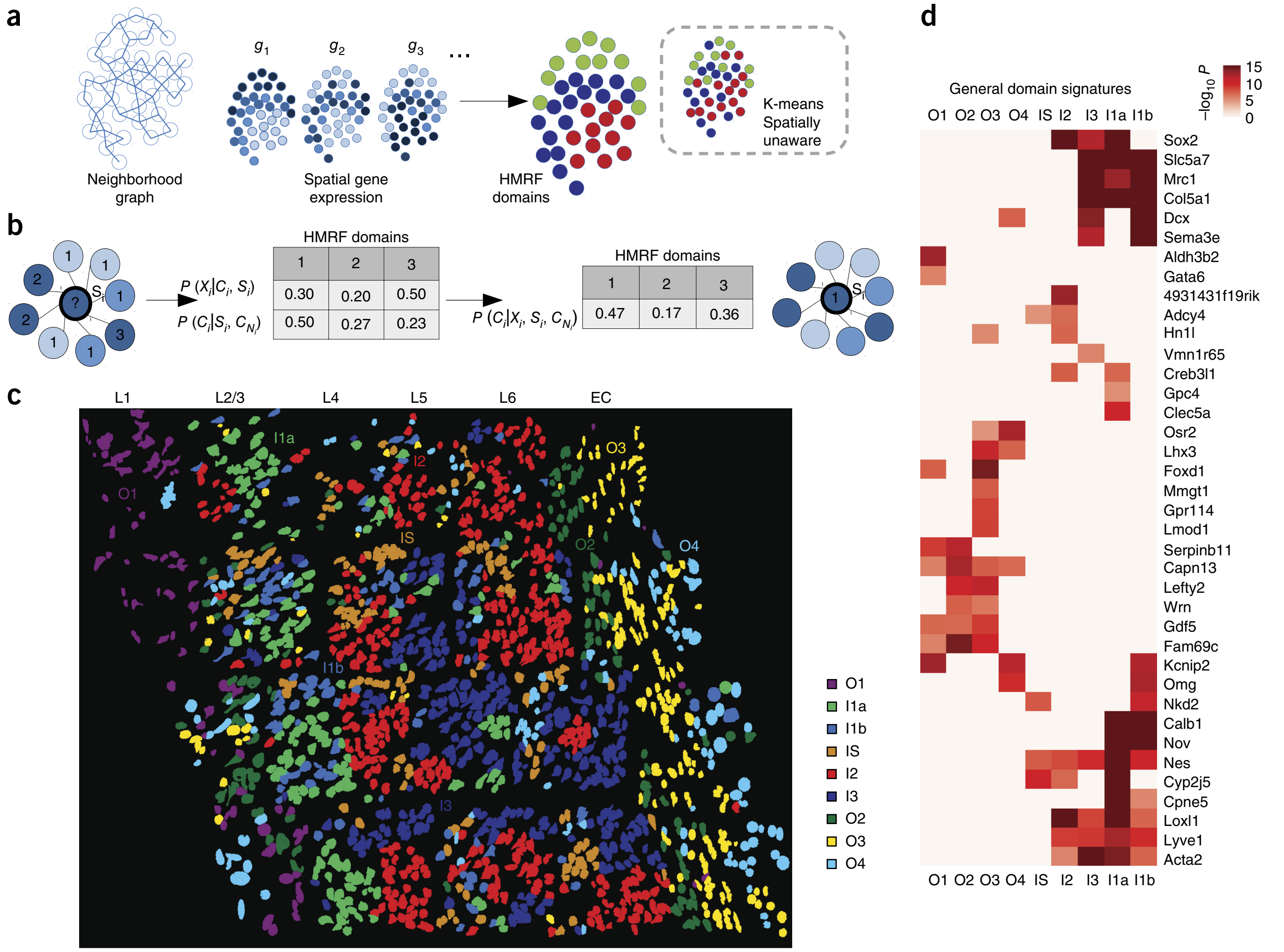

Figure 2 Spatial domain dissection in seqFISH data using HMRF. (**a**) A schematic overview of the HMRF model. A neighborhood graph represents the spatial relationship between imaged cells (indicated by the circles) in the seqFISH data. The edges connect neighboring cells. seqFISH-detected multigene expression profiles are used together with the graph topology to identify spatial domains. In contrast, k-means and other clustering methods do not utilize spatial information and the results are therefore expected to be less coherent (illustrated in the dashed box). (**b**) An intuitive illustration of the basic principles in a HMRF model. For a hypothetical cell (indicated by the question mark), its spatial domain assignment is inferred from combining information from gene expression (x_i) and neighborhood configuration (c_{N_i}). The color of each node represents the cell's expression and the number inside each node is the domain number. In this hypothetical example, combining such information results in the cell being assigned to domain 1, instead of domain 3 (Online Methods). (**c**) HMRF identified spatial domain configuration in the mouse visual cortex region. Distinct domains revealed a resemblance to layer organization of cortex. Naming of domains: I1a, I1b, I2 and I3 are inner domains distributed in the inner layers. O1–O4 are outer domains. IS represents the inner scattered state. These domains are associated with cell morphological features, such as distinct cell shape differences in outer layer domains. Cell shape information was obtained from the segmentation of cells from images (Online Methods). For HMRF, 1,000 initial centroids were used and the best configuration was selected to initialize HMRF. The procedure was repeated two more times with similar results. (**d**) General domain signatures are shared between cells within domains. *P* values signify two-sided Welch's *t* tests with *P* values adjusted for multiple comparisons. Genes with significant *P* values are shown. All domains are compared: O2 (n = 109 cells), I1a (n = 389), O4 (n = 120), I1b (n = 79), O1 (n = 135), I2 (n = 117), I3 (n = 205), O3 (n = 270) and IS (n = 173).

(<10 copies per cell). Cross-validation analysis revealed that, using these 43 genes as input, the SVM model accurately mapped 90.1% of the cells in the scRNAseq data to the correct cell type. Thus, we used these 43 genes (**Supplementary Table 2**) to map cell types in the seqFISH data.

As a first step, we preprocessed the seqFISH data using a multi-image regression algorithm to reduce potential technical biases resulting from non-uniform imaging intensity variation (Online Methods). We further adopted a quantile normalization[30] approach to calibrate the scaling and distribution differences between scRNAseq and seqFISH experiments. For most genes, the quantile-quantile (q-q) plot normalization curve was notably linear (**Supplementary Fig. 1**), suggesting a high degree of agreement between the two data sets despite technological differences. We then applied the SVM classification model to the bias-corrected, quantile-normalized seqFISH data to assign cell types. Of note, we found that better performance could be achieved by further calibrating model parameters to accommodate platform differences. The results of multiclass SVM were calibrated across models[31] and converted to probabilities. We found that 5.5% cells were excluded, that is, they could not be confidently mapped to

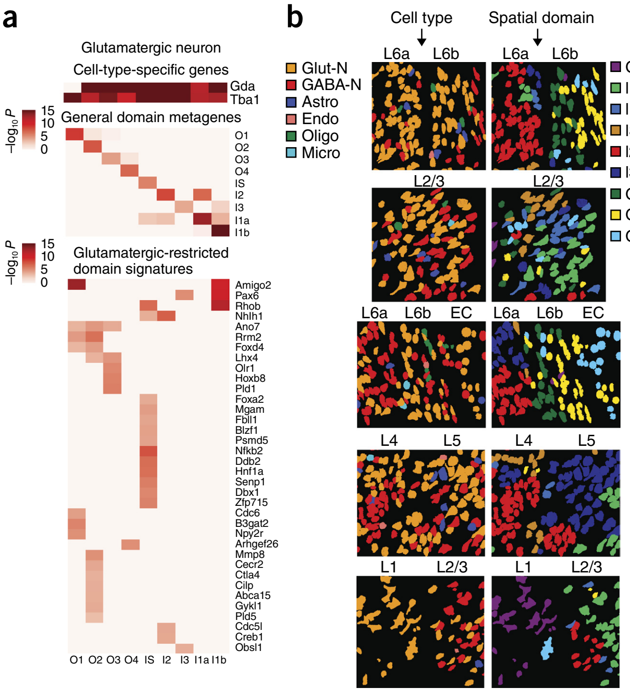

Figure 3 HMRF analysis identified domain-associated heterogeneity in glutamatergic cells. (**a**) Three major sources of variations in glutamatergic neurons ($n = 859$). Glutamatergic neurons were distributed across nine domains with 79, 187, 88, 58, 93, 60, 73, 129 and 92 cells in O2, I1a, O4, I1b, O1, I2, I3, O3 and IS domains, respectively. Top, cell-type-specific signals: *Gda* and *Tbr1*. Middle, general domain signatures, as shown in **Figure 2d**, were summarized as metagene expression. Bottom, glutamatergic-restricted domain signatures, as identified by comparing glutamatergic cells across domains and removing signatures that were general domain signatures. Signature genes were obtained by two-sided Welch's *t* tests with *P* values adjusted for multiple comparisons. (**b**) Snapshots of single cells. Each row shows a snapshot of cells at the boundary of two layers. Each of two columns is a type of annotation. Left, cell type; right, HMRF domains. Cell type annotation was incapable of explaining layer-to-layer morphological variations; for example, glutamatergic cells (orange) were present in all layers, yet morphological differences existed in glutamatergic cells. HMRF domains better captured the boundary of two layers in each case, in that the domains could separate distinct morphologies. A systematic comparison is shown in **Supplementary Figure 12** (see also **Supplementary Fig. 11**).

a single cell type (with 0.5 or less probability). Among the mapped cells, 54% were glutamatergic neurons, 37% were GABAergic neurons, 4.8% were astrocytes, and other glial cell types and endothelial cells comprised the remaining 4.2% of cells (**Fig. 1c**).

To validate our predictions, we first checked the expression of known marker genes and compared the average gene expression profiles between scRNAseq and seqFISH data. Indeed, this comparison revealed a high degree of similarity (**Fig. 1c**). Notably, marker genes were, as expected, highly expressed in the matched cell types, such as *Gja1* and *Mfge8* in astrocytes, *Laptm5* and *Abca9* in microglia, *Cldn5* in endothelial cells, *Tbr1* and *Gda* in glutamatergic neurons, and *Slc5a7* and *Sox2* in GABAergic neurons. The majority of cell types had a high Pearson correlation (>0.8) between matched cell types' average expression profile; even for the rare cell type microglia, the correlation remained reasonably high (0.75) (**Fig. 1d**). We were also able to distinguish early maturing oligodendrocytes in the seqFISH data on the basis of *Itpr2* expression (**Fig. 1c**), as previously reported[15]. Expression patterns of inhibitory GABAergic neurons and excitatory glutamatergic neurons exhibited strong anti-correlation (**Fig. 1d**).

As an additional validation, we compared the neurons that were stained with Nissl and DAPI with astrocytes that were only stained with DAPI. Our cell type mapping results agree with these patterns. Over 89% of predicted astrocytes exhibited strong DAPI staining, but weak or no Nissl staining, across cortex columns (**Supplementary Note 1** and **Supplementary Table 3**). Taken together, these analyses indicate that the majority of cells were mapped to the correct cell types.

By combining cell type predictions from scRNAseq and positional information from seqFISH, we were able to construct a single-cell resolution landscape of cell type spatial distribution (**Fig. 1e**). As expected, this landscape is very complex, with different cell types intermixed with each other (**Fig. 1e**). On the other hand, it is clear that there remains substantial heterogeneity in each cell type.

A systematic approach to identify multicellular niche

Microenvironment in tissues can contribute to heterogeneity in addition to cell-type-specific expression patterns. To systematically dissect the contributions of microenvironments on gene expression variation, we developed a hidden-Markov random field (HMRF) approach[32] to unbiasedly inform the organizational structure of the visual cortex (**Fig. 2a**). The basic assumption is that the visual cortex can be divided into domains with coherent gene expression patterns. A domain may be formed by a cluster of cells from the same cell type, but it may also consist of multiple cell types. In the latter scenario, the expression patterns of cell-type-specific genes may not be spatially coherent, but environment-associated genes would be expressed in spatial domains. A HMRF enables the detection of spatial domains by systematically comparing the gene signature of each cell with its surroundings to search for coherent patterns. Briefly, we computationally constructed an undirected graph to represent the spatial relationship among the cells, connecting any pair of cells that were immediate neighbors (**Fig. 2a,b**). Each cell was represented as a node in this graph. The domain state of each cell was influenced by two sources (**Fig. 2b**): its gene expression pattern and the domain states of neighboring cells. The total contribution of neighboring cells can be mathematically represented as a continuous energy field, and the optimal solution is identified by searching for the equilibrium of the field (Online Methods and **Supplementary Note 1**).

Next, we applied our HMRF model to analyze the 1,597-cell mouse visual cortex seqFISH data set. The expression of the 125 genes ranged from being highly scattered to spatially organized. To enhance spatial domain detection, we defined a spatial coherence score and selected the top 80 genes for HMRF analysis (Online Methods). As an additional filter, we further removed 11 genes that were highly specific to a single cell type, resulting in 69 genes (**Supplementary Table 4**) for spatial domain identification. We found that this additional filtering step improved the resolution while preserving the overall spatial pattern (**Supplementary Fig. 2**).

HMRF modeling of the visual cortex region revealed nine spatial domains (**Fig. 2c**). These domains had distinct spatial patterns; some displayed a layered organization that resembled the anatomical structure[33]. For example, four of the domains were located on the outer layers of the cortex, and we labeled them as O1, O2, O3 and O4 (**Fig. 2c**). The locations of these layers roughly corresponded to the well-characterized L1, L6 and external capsule (EC) layers, respectively. Four domains were located on the inside of the cortex therefore labeled as I1a, I1b, I2, and I3, respectively (**Fig. 2c**). These domains roughly corresponded to the L2–5 layers. These inner domains were less pronounced than the outer domains, which is consistent with previous

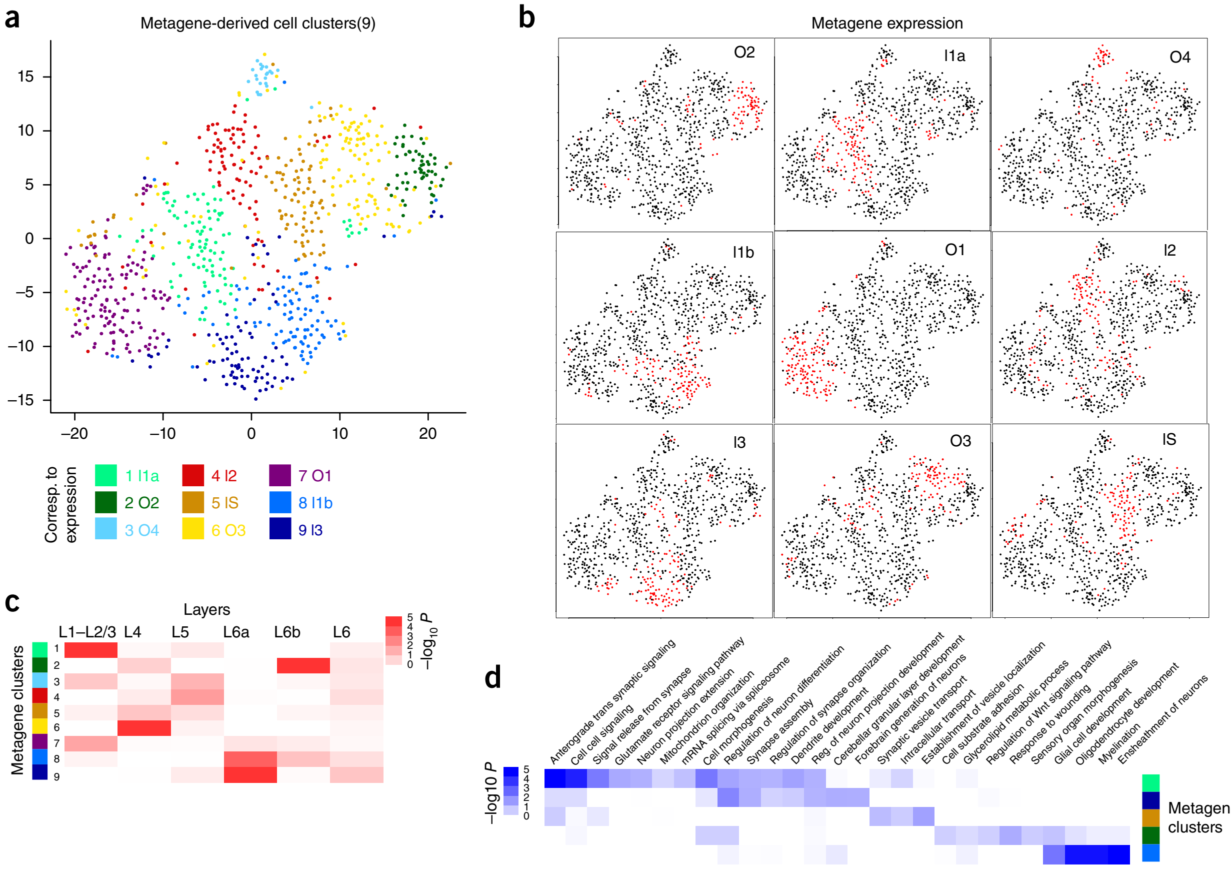

Figure 4 Reanalysis of single-cell RNAseq data (from ref. 27) with domain signatures summarized into metagenes. (**a**) t-SNE plot shows how the 812 glutamatergic cells from ref. 27 cluster according to expanded domain signatures aggregated as metagenes (shown in **b**). Colors indicate k-means clusters (k = 9). Each cluster is annotated by its enriched metagene expression. Nine annotated glutamatergic metagene clusters were identified: O2 (n = 132 cells), I1a (n = 98), O4 (n = 92), I1b (n = 131), O1 (n = 84), I2 (n = 22), I3 (n = 97), O3 (n = 100) and IS (n = 56). (**b**) Binarized metagene expression profiles for the glutamatergic cells. Red, population that highly expresses the metagene. (**c**) Spatial clusters defined according to metagenes were enriched in manual layer dissection annotations. Column, layer annotation information obtained from microdissection[27], with L1–L2/3 (n = 48 cells), L4 (n = 202), L5 (n = 116), L6 (n = 12), L6a (n = 87) and L6b (n = 33). Row, metagene-based cell clusters. Shown are hypergeometric P values of cell overlaps. (**d**) Inferred spatial clusters of glutamatergic neurons were enriched in distinct GO biological processes. Shown are hypergeometric P values of gene overlaps between differentially expressed genes (n = 500) of each metagene cluster and Gene Ontology gene sets (variable sizes). P values were adjusted for multiple comparisons.

anatomical analysis. Finally, one domain was sporadically distributed across in the inner layers of the cortex, and we labeled it as IS (**Fig. 2c**). Of note, such domain-like patterns were not visible in the cell type localization pattern (**Fig. 1e**). Consistent with these results, a t-distributed stochastic neighbor embedding (t-SNE) plot using these 69 genes identified clustering patterns that were similar to the domain annotations but differed greatly from the cell type annotations (**Supplementary Fig. 3**). These results strongly suggest that HMRF provides complementary information to cell type annotations.

By overlaying cell type annotations, we found that each domain generally consisted of a mixture of GABAergic neurons, glutamatergic neurons and astrocytes interacting in each environment (**Supplementary Fig. 4**). The decomposition of mouse visual cortex into spatial domains suggests that a spatial gene expression program is shared across cells in proximity. Differential gene expression analysis identified distinct signatures, which we labeled as the general domain signatures, associated with each spatial domain (**Fig. 2d** and **Supplementary Figs. 5–7**). For example, the genes *Calb1*, *Cpne5* and *Nov* were preferentially expressed in inner domains (I1a, I1b), whereas *Serpinb11* and *Capn13* were highly enriched in outer domains (O1, O2). Different outer domains could be further distinguished by additional markers, such as *Mmgt1* (O3), *Aldh3b2* (O1) and *Fam69c* (O2). Notably, these spatial gene signatures transcended multiple cell types and were therefore distinct from cell-type-specific signatures (**Supplementary Figs. 6** and **7**). The spatial marker genes, including *Calb1*, *Cpne5* and *Nov*, were highly consistent with their spatial expression in Allen Brain Atlas[33] ISH images (**Supplementary Fig. 8**). Other markers, such as *Nell1*, *Aldh3b2* and *Gdf5*, had layer-specific expressions that were consistent with previous results[15] (**Supplementary Fig. 8**). We summarized the gene signature of each domain as a metagene, defined as the average expression of the subset of genes that were specifically associated with the domain. This provides an 'analog' representation of the spatial domain information as an additional diagnostic (**Supplementary Fig. 9**). Taken together, these analyses strongly suggest that our model for analyzing seqFISH data is able to detect functionally and transcriptionally distinct spatial environments.

Interactions between cell type and spatial environment

Glutamatergic neurons mediate the neuronal circuit in the visual cortex via a primarily excitatory function. It is also well-known that

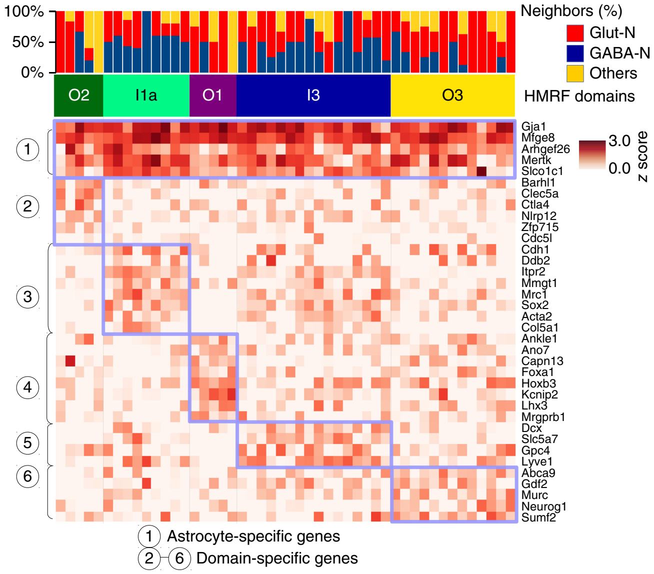

Figure 5 Spatially dependent astrocyte variation revealed by HMRF. Neighborhood cell type composition for the 48 astrocyte cells (columns). Cells are ordered by HMRF domain annotations. The heat map shows single-cell expression of astrocytes clustered by domain-specific genes. Blue-box highlights the common signatures expressed in each domain's astrocyte population. Identified astrocyte subpopulations are O2 (n = 5), I1a (n = 9), O1 (n = 5), I3 (n = 16) and O3 (n = 13).

the behavior of different glutamatergic neurons can be very different[27,34]. Using a combination of cell type mapping and spatial domain identification, we set out to dissect the source of heterogeneity in glutamatergic cells. First, nearly all glutamatergic cells expressed cell-type-specific markers, such as *Gda* and *Tbr1* (**Fig. 3a**). In addition to cell type identity, there exists substantial spatially dependent heterogeneity in glutamatergic cells. Given that glutamatergic cells are spread across all nine domains, each subset expressed a different gene signature in accordance with domain annotation (**Fig. 3a**). First, the general domain signatures (**Fig. 2d**), aggregated as metagenes, could be used to separate glutamatergic cells into domains (**Fig. 3a**). Second, beyond the general signature, an additional set of gene signatures were differentially expressed between glutamatergic cells in different domains (**Fig. 3a**). To distinguish these genes from the general domain signatures, which transcend cell types, we referred to these genes as the glutamatergic-restricted signatures. For example, *Mmp8* expression was restricted to domain O2 (**Fig. 3a**), whereas *Hoxb8* expression was specific to O3 and *Nfkb2* to IS (**Fig. 3a**). Collectively, the domain-specific signatures mapped out the spatial patterns of expression in glutamatergic cells, demonstrating their power to differentiate subgroups of this cell type (**Supplementary Figs. 9–11**).

By visual inspection, we observed notable morphological variations near the boundary between different domains at multiple regions (**Fig. 3b**), including change of circularity and cell orientations, and these were accompanied by metagene expression switches (**Supplementary Fig. 11**). To systematically compare the morphological differences between different domains, we extracted quantitative information of 15 different morphological features per cell based on the Nissl staining images and compared the statistical distributions across different domains. Indeed, we found that a number of features displayed strong domain associations, including circularity in O4 ($P < 6.1 \times 10^{-12}$), width in I1b ($P < 1.6 \times 10^{-14}$), angle in O3 ($P < 6.7 \times 10^{-18}$) and minimum feret diameter in I1a ($P < 3.0 \times 10^{-11}$) (**Supplementary Fig. 12**). Of note, these differences could not be identified by cell type mapping alone (**Fig. 3b**). Thus, in neuronal cell types, such as glutamatergic or GABAergic neurons, substantial morphological differences remain across domains, suggesting that spatial positions account for a large part of the morphologies of these cells, consistent with known morphological diversity in the cortex. Overall, these analyses strongly suggest that spatial domain variation is important for mediating cellular heterogeneity.

Using HMRF domain information to reanalyze scRNAseq data

ScRNAseq data does not contain spatial information. However, using domain signatures derived from seqFISH as a guide, we were able to infer spatial locations from scRNAseq data. To dissect the contribution of environmental factors to transcriptomic heterogeneity, we focused on glutamatergic cells and combined the general domain signatures with the additional set of markers that are glutamatergic restrictive. Using these expanded domain signatures (**Supplementary Table 5**) summarized as metagenes, we were able to uncover a hidden structure in the glutamatergic cells (**Fig. 4a**,**b**). Notably, the glutamatergic cells could be partitioned into nine different clusters on the basis of the expanded domain signatures, which were highly consistent with seqFISH data analysis (**Fig. 4a**,**b**). As such, these clusters were labeled according to their enriched metagene signatures (**Fig. 4a**).

We compared the inferred domain annotations with the original sites of dissection in a previous study[27]. Several domains matched the corresponding layer structure very well (**Fig. 4c**). For example, cluster 1 (annotated as domain I1a based on metagene analysis) significantly overlapped with L1–L2/3 ($P < 2.3 \times 10^{-6}$). Cluster 2 (annotated as domain O2) overlapped with L6b ($P < 4.8 \times 10^{-9}$), and cluster 9 (annotated as domain I3) significantly overlapped with L6a ($P < 1.0 \times 10^{-8}$). On the other hand, clusters 3–5 (annotated as domains O4, I2 and IS) did not correspond to specific layers.

Using the whole transcriptomes from scRNAseq, we searched for additional domain-specific gene signatures based on coexpression analysis. Our analysis identified a number of genes that were not assayed by seqFISH, including *Tubb2a* (I1a) and *Ndrg3* (O4). We examined the corresponding ISH images in the Allen Brain Atlas and found that the inferred spatial patterns agreed well with the imaging data (**Supplementary Fig. 13**). We further conducted gene set enrichment analysis based on the inferred domain-specific markers and identified a number of functional biological processes that were enriched in specific domains (**Fig. 4d**).

An important question is whether the distinction between the subpopulations identified through our integrative analysis simply reflect cell subtype differences that can be identified through scRNAseq analysis alone. To address this question, we systematically compared the domain and cell subtype annotations using a number of approaches, including the underlying gene signatures, the grouping of cells based on domain or cell subtype annotations, and tSNE-based visualizations (**Supplementary Figs. 14** and **15**). Based on these comparisons, we came to two conclusions. On one hand, we observed a non-negligible association between the two sets of annotations, such as at L6b_Serpinb11, L2/3_Ptgs2 and L6a_Sla (**Supplementary Fig. 14**). For example, several domain-specific markers were also markers of specific cell subtypes, such as *Serpinb11*, *Cpne5* and *Sema3e* (**Supplementary Fig. 16a**). On the other hand, it was also clear that the overall structure of domain and subtype annotations were very different. For example, cells whose locations we inferred to be in domains O1, IS and O4 spread across multiple subtypes

(**Supplementary Figs. 14** and **16b**). Conversely, neither the L5a_Batf3 nor L5a_Hsd11b1 subtype was associated with any specific domain (**Supplementary Fig. 14**). Taken together, these analyses strongly indicate that the domain patterns are distinct from, and therefore complementary to, cell subtype annotations. Thus, integrating seqFISH data analysis provides new insights into scRNAseq data.

Region-specific variation among astrocytes

Next, we investigated the environment effect on astrocytes, which are also known to have substantial heterogeneity[20,35]. Our cell type mapping identified 48 astrocytes in the seqFISH data. These cells all expressed key astrocyte markers but were spread across five different spatial domains (O1, O2, O3, I1a and I3; **Fig. 5a**). Of note, a number of astrocyte markers[20] were only expressed in specific domains (**Supplementary Fig. 17**). As an example, *Acta2*, *Col5a1* and *Sox2* were strongly associated with domain I1a, whereas their expression levels were greatly reduced in domains O1 and O2. On the other hand, the expression levels of *Clec5a* and *Ankle1* were high in domains O2 and O1 but were much lower in other domains. The spatially dependent variations might underline important functional differences.

DISCUSSION

A major goal in single-cell analysis is to systematically dissect the contributions of cell types and environment to cell-state variability. We developed an HMRF-based computational approach to combine the strengths of sequencing and imaging-based single-cell transcriptomic profiling strategies. We used our method to detect spatial domains in the mouse visual cortex region. In doing so, we were able to identify environment-associated variations. Our analysis also demonstrated that further insights can be gleaned from single-cell data by integrating information from complementary technologies. In particular, integrating scRNAseq data allowed us to map cell types more accurately than using seqFISH data analysis alone, whereas integrating seqFISH data allowed us to extract spatial structure in scRNAseq data analysis. Although the classification of a small number of isolated cells as domains may be questionable, such events were rare and did not affect the overall spatial domain patterns.

To test the generalizability of our method, we used it to analyze a published spatial transcriptomic data set obtained from a very different technology at olfactory bulbs[36]. Here, spatial information was identified by hybridizing mRNA to a specially designed tissue microarray containing spatial barcoding oligo-probes. Despite the substantial platform differences, our HMRF model was able to recapitulate the spatial domains that are consistent with the underlying anatomical structures (**Supplementary Fig. 18**). In another example, we analyzed seqFISH data[19] obtained from a different region of the mouse brain (dentate gyrus) using different probes. Again, the results were consistent with the anatomical structure (**Supplementary Fig. 19**). These analyses strongly indicate that our method is generally applicable. Of note, our HMRF model is agnostic about the cell type composition and associated gene signatures. Moreover, its application does not require single-cell resolution data, as it can also detect spatial patterns on larger scales.

Two recent studies have also investigated spatially variable genes. Specifically, SpatialDE[37] is designed to identify individual genes whose expression levels at neighboring sites are correlated. Of note, SpatialDE does not identify spatial regions with distinct expression patterns. Similarly, trendsceek[38] is also designed to detect spatial dependency. However, its application is limited to a single gene at a time. In contrast, our HMRF method can simultaneously detect the combinatorial pattern of all profiled genes. A unique aspect of our study was the integration of cell type and spatial domain annotations. This is important for systematically dissecting the roles of intrinsic regulatory networks and spatial environment in the maintenance of cellular states. Future work will be needed to investigate the mechanisms underlying the interactions between cell type and microenvironment.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported by a Claudia Barr Award, a Chan Zuckerberg Initiative Award, and NIH grant R01HL119099 to G.-C.Y., and by grants from the Paul G. Allen Foundation Discovery Center, NIH HD075605 and TR01 OD024686 to L.C.

AUTHOR CONTRIBUTIONS

G.-C.Y. and L.C. conceived and supervised the project. Q.Z. and G.-C.Y. conceived the HMRF and SVM models. Q.Z. and G.-C.Y. conducted and supervised the computational analyses. S.S. and L.C. conducted and supervised the seqFISH experiments. Q.Z., S.S., R.D., G.-C.Y. and L.C. wrote the manuscript. All of the authors contributed ideas for this work. All of the authors reviewed and approved the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Quail, D.F.D. & Joyce, J.A. Microenvironmental regulation of tumor progression and metastasis. *Nat. Med.* **19**, 1423–1437 (2013).
2. Riquelme, P.A., Drapeau, E. & Doetsch, F. Brain micro-ecologies: neural stem cell niches in the adult mammalian brain. *Phil. Trans. R. Soc. Lond. B* **363**, 123–137 (2008).
3. Swain, P.S., Elowitz, M.B. & Siggia, E.D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 12795–12800 (2002).
4. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
5. Zhang, J. & Li, L. Stem cell niche: microenvironment and beyond. *J. Biol. Chem.* **283**, 9499–9503 (2008).
6. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).
7. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
8. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
9. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
10. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
11. Jaitin, D.A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
12. Kolodziejczyk, A.A. *et al.* Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
13. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
14. Shekhar, K. *et al.* Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).
15. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
16. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
17. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012).
18. Moffitt, J.R. *et al.* High-performance multiplexed fluorescence *in situ* hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. USA* **113**, 14456–14461 (2016).
19. 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 (2016).

20. Zhang, Y. *et al.* Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
21. Yuan, G.C. *et al.* Challenges and emerging directions in single-cell analysis. *Genome Biol.* **18**, 84 (2017).
22. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
23. Halpern, K.B. *et al.* Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature* **542**, 352–356 (2017).
24. Karaiskos, N. *et al.* The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
25. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
26. Joost, S. *et al.* Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Syst.* **3**, 221–237 (2016).
27. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
28. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
29. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
30. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
31. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**, 61–74 (1999).
32. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).
33. Sunkin, S.M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).
34. Andjelic, S. *et al.* Glutamatergic nonpyramidal neurons from neocortical layer VI and their comparison with pyramidal and spiny stellate neurons. *J. Neurophysiol.* **101**, 641–654 (2009).
35. Ben Haim, L. & Rowitch, D.H. Functional diversity of astrocytes in neural circuit regulation. *Nat. Rev. Neurosci.* **18**, 31–41 (2017).
36. Ståhl, P.L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
37. Svensson, V., Teichmann, S.A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).
38. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).

ONLINE METHODS

SeqFISH data generation. SeqFISH data in the mouse visual cortex region was generated as described previously[19]. Briefly, 100 genes were encoded using a temporal barcoding method and 25 genes were quantified individually. To encode 100 genes, four rounds of hybridization were performed using five distinct fluorescence channels. Out of a total possible 625 barcodes, 100 were chosen such that loss of signal in any given hybridization still allows accurate decoding of the spot. Every transcript was hybridized in every round using a given probe set. After hybridization, the signal was amplified using smHCR and images were taken at predefined locations in the mouse visual cortex. The DNA probes along with the amplification polymers were digested using DNase I leaving behind a naked RNA for re-hybridization with the next probe set. A round of imaging with DAPI staining (which labels the DNA) was done before any RNA hybridization to image all nuclei in the fields and a final round of Nissl staining (which labels the cell body in neuronal cells) was imaged to identify cell boundaries. Cells were segmented based on DAPI staining, Nissl staining, and RNA point density. Once all imaging rounds were completed, these images were aligned using a two-dimensional (2D) normalized cross correlation and each spot was decoded based on the unique color switching pattern. For the 25 genes that were labeled without any barcoding, simple spot counting was done to identify the number of transcripts. These transcripts were then assigned to cells based on the location of the transcript and the segmentation masks. For more details regarding the seqFISH method, please refer to ref. 19. The spatial coordinates of the cells are provided on our website (see data availability).

SeqFISH data normalization and bias correction. The seqFISH gene expression matrix, represented by log(count + 1), was normalized by row and column z-scoring to remove cell-specific and gene-specific biases. Potential field imaging biases were estimated and removed by using a multi-image regression algorithm similar as previously done[39]. Briefly, for each gene, the imaging bias at each binned location was estimated by averaging the normalized gene expression levels over eight neighboring bins in each field followed by averaging across all fields. The estimated bias was then modeled by principal component analysis. The contributions of the four most significant principal components were estimated by linear regression and removed from the normalized gene expression matrix (**Supplementary Fig. 20**).

Cell type mapping. Single-cell RNAseq data for the mouse visual cortex were obtained from Gene Expression Omnibus[40] (GSE71585). Cell type information corresponding to 1,723 cells was obtained from the original paper[27]. In this analysis, we considered the eight major cell types: GABAergic, glutamatergic, astrocytes, three oligodendrocyte groups, microglia, and endothelial cells. Differentially expressed genes among different cell types were identified by MAST[41].

We trained classifiers of cell types from single-cell RNAseq data set by using the multiclass SVM formulation. For each cell type, we built a classifier as follows. Let x_i, $i = 1,\ldots, n$, be the gene expression pattern for the i-th cell, and y_i code for cell type identity: $y_i = 1$ if cell i belongs to the specified cell type and -1 otherwise. We selected the linear kernel that produces two hyperplanes that best separates the two classes. The objective function is defined as follows

$$\text{minimize } C\left(\sum_{i=1}^{n}\zeta_i^2\right)+\|w\|^2/2$$

$$\text{subject to } 1-\zeta_i \le y_i(w\cdot x_i-b),\ \zeta_i \ge 0 \qquad (1)$$

Here w is the normal vector to the hyperplane used to represent margin. The squared hinge loss function $\Sigma_{i=1}^{n}\zeta_i^2$ is used here to quantify the margin of misclassification error. C is a regularization parameter that trades off misclassification due to overfitting against simplicity of the decision function. A lower C increases the ability of the model to generalize to unseen data at a cost of larger fitting error. For testing data, the sign of $w{\cdot}x_i - b$ is used to predict cell type identity. We used the Python LinearSVC implementation, which is part of the scikit-learn 0.19 library[42], with the following parameter setting: class_weights = balanced, dual = False, max_iter = 10,000, and tol = 1×10^{-4}.

Using the SVM model formulated as above, we first tested how many genes are needed for accurate cell-type mapping. To this end, we randomly subset 20, 40, 60 and 80 genes from the list of differentially expressed genes and, for each gene set, built a vanilla SVM classification model to map each cell in the single-cell RNAseq data set to its corresponding cell type. The accuracy was evaluated by using fourfold cross-validation. Our results indicated that a high accuracy (>90%) can be obtained with 40 or more genes.

In addition to the major cell types mentioned above, a previous study[27] also identified 22 fine cell classes, and 49 minor cell classes. Using the same approach, we also evaluated the accuracy of refined cell type mapping (**Supplementary Fig. 21**). We found that approximately 200 genes were required to achieve 85% accuracy in predicting 22 finer classes, and over 800 genes were needed to predict the 49 minor cell types with 75% accuracy. Therefore, we focused on the mapping of eight major cell types on seqFISH given that they can be predicted accurately with fewer than 100 genes (ROC curves in **Supplementary Fig. 22**).

To map cell types in the seqFISH data, we made a few modifications to incorporate the platform differences. First, since 125 genes were profiled by seqFISH, we used the intersection with the top differentially expressed genes ($P < 1 \times 10^{-20}$) in the scRNAseq data set for cell type mapping, thereby selecting 43 genes in total. Based on the subsampling analysis described above, these 43 genes were sufficient for accurate cell type mapping. Second, the scRNAseq data were z-score transformed so that the dynamic range was comparable with seqFISH data. Third, we used quantile normalization[30] to convert seqFISH data so that the statistical distribution was almost identical to single-cell RNAseq data. Fourth, we chose the regularization parameter C to maximize the cross-platform correlation between the cell-type specific gene expression profiles, resulting an estimate of $C = 1 \times 10^{-6}$. Finally, to account for the possibility that certain cells cannot be unequivocally assigned to a single cell type, we used Platt scaling[31] to convert SVM output to a probability measure and then selected a cutoff value of 0.5 probability to filter cells that can be confidently mapped to a single cell type. 97 (5%) cells did not pass this filter.

HMRF. HMRF is a graph-based model commonly used for pattern recognition in image data analyses[32,43]. In a common setting, HMRF is used to model the spatial distribution of a signal, such as the pixel intensities over a 2D image. The spatial structure is represented as a set of nodes on a regular grid, where neighboring nodes are connected to each other. The spatial pattern is 'hidden' in the sense that it must be indirectly estimated from other variables that can be directly measured. The most important assumption in HMRF is the Markov property, which states that the spatial constraints can be reduced to considering only correlation between immediate neighboring nodes. This simplifying assumption implies that the joint distribution can be decomposed as products of much smaller components each defined on a fully connected subgraph (termed cliques). As has been done previously, we decomposed the graph into size-2 components (or edges in the graph) that provides a convenient means to estimating the MRF by using pairwise energies.

Specifically, let $S = \{S_i\}$ be the nodes in the graph. The set of nodes and the adjacency relation as defined by the local neighborhood graph forms the neighborhood system $(S, \{N_i\})$. Every node is associated with observed signal values x_i. Let $C = \{c_i = 1,\ldots,K\}$ represent the set of possible classes of patterns. The joint probability that a node S_i is associated with class c_i is specified by the following equation:

$$P(c_i\,|\,s_i,x_i,c_{N_i})=1/Z\,P(x_i\,|\,c_i,s_i)P(c_i\,|\,s_i,c_{N_i}) \qquad (2)$$

In the right hand side, the term $P(x_i|c_i,s_i)$ models the effect of the node S_i's own gene expression, whereas $P(c_i\,|\,s_i,c_{N_i})$ models the effect of the neighboring cells configuration c_{N_i}. The combined effect of these two terms is schematically shown in **Figure 2**. The latter term is further determined by the Gibbs distribution:

$$P(c_i\,|\,s_i,c_{N_i})=1/Z_2\exp\left(-\beta\sum_{s_j\in N_i}U(c_j,c_i)\right) \qquad (3)$$

where $U(c_j, c_i)$ is referred to as the energy function. The exact formulation of $U(c_j, c_i)$ is dependent on the specific application, and it imposes the assumption of how neighboring nodes interact with each other. Here we use the special case Pott's model.

$$U(c_j,c_i)=-1 \text{ if } c_j=c_i \text{ and } 0 \text{ otherwise} \qquad (4)$$

which means that the effects of neighboring cells are additive. Essentially, $P(c_i \mid s_i, c_{N_i})$ expresses the total energies as a summation of pairwise interaction energies with neighbors. The parameter beta reflects the strength of interactions.

Application to seqFISH data. The HMRF model described above is naturally applicable to analyze seqFISH data. Here each class of patterns corresponds to a spatial domain. The observed signals are gene expression levels measured by seqFISH data, whose distribution is modeled as a multivariate Gaussian random variable. The application of HMRF to seqFISH data analysis involves the following four components: neighboring graph representation, gene selection, domain number selection, and model inference. The details of each component are described below.

1. Neighborhood graph representation. An undirected graph was constructed to represent the spatial relationship between the cells. Each node represents a cell, and each edge connects a pair of neighboring cells. The neighborhood size was chosen such that on average each cell has five neighboring cells.

2. Gene selection. We selected a subset of genes whose expression patterns tend to be spatially coherent based on the following analysis. For each gene *g*, cells were divided into two mutually exclusive sets: the first set, denoted by L_1, contains cells with high expression at the 90th percentile expression level cutoff, and the rest of the cells were denoted by L_0. The spatial coherence of gene expression was quantified as the Silhouette coefficient[44] of the spatial distance associated with these two cell sets. Specifically, the Silhouette coefficient is calculated as:

$$\delta_g = 1/|L_1| \sum_{s_i \in L_1} (m_i - n_i)/\max(m_i, n_i) \qquad (5)$$

where for a given cell s_i in set L_1, m_i is defined as the average distance between s_i and any cell in L_0, and n_i is defined as the average distance between s_i and any other cell in L_1. Here, we used the rank-normalized, exponentially transformed distance to quantify the local physical distance between two cells. For a pair of cells s_i and s_j, this distance is defined as $r(s_i, s_j) = 1 - q^{\mathrm{rank}_d(s_i, s_i) - 1}$ where $\mathrm{rank}_d(s_i, s_j)$ is the mutual rank[45] of s_i and s_j in the vectors of euclidean distances $\{\mathrm{Euc}(s_i, *)\}$ and $\{\mathrm{Euc}(s_j, *)\}$. Hence, this exponentially weighted function[46] is designed to place more emphasis on closely located cells and penalize far-away cells' distance. *q* is a rank-weighting constant ($0 < q < 1.0$) set at 0.95. The statistical significance of δ_g was evaluated by random permutation, and the genes associated with significant values of δ_g ($P < 0.05$) were selected as spatially coherent.

Using the above criteria, we found 80 spatially coherent genes. We further removed 11 cell type specific genes (MAST $P < 1 \times 10^{-20}$) which have average expression z-score > 2. We found this additional filtering step is useful for improving the resolution while preserving the overall spatial pattern (**Supplementary Fig. 2**). We repeated the analysis using varying degree of stringency for selecting spatially coherent genes (**Supplementary Fig. 23**), varying the degree of excluding cell-type-specific genes (**Supplementary Fig. 2**), and varying beta (**Supplementary Fig. 24**), and found that the overall patterns identified by the HMRF model is robust against these variations.

3. Domain number selection. We used *k*-means clustering results as initialization for the HMRF domains. The value of *k* was selected based on the gap-statistics[47].

4. Implementation and model inference. The model parameters were inferred by using the Expectation-Maximization (EM) algorithm[48]. We developed a new implementation based on the MRITC R package[49] and GraphColoring Java package[50]. The implementation contains modifications to accommodate arbitrary neighborhood graph topology. The domain assignment for each cell was determined by using *maximum a posteriori* estimation, which can be viewed as the equilibrium state of the energy function. See **Supplementary Note 1** for implementation details.

Robustness analysis of the HMRF model. We also tested the robustness of our HMRF model against spatial perturbation. This was achieved by randomly exchanging the spatial locations of a subset of cells (10, 20, 40 and 100%). At 100% exchanging rate, the spatial coherence is completely disrupted. Log-likelihood of the HMRF model was recorded and compared across scenarios. As expected, the log-likelihood achieves maximum at a low perturbation rate and gradually decreases as the exchange rate increases. The difference between the perturbed and unperturbed data is highly statistically significant (**Supplementary Fig. 25**).

Domain-specific gene signatures. For each spatial domain, we identified a subset of genes that were significantly upregulated in the domain compared to cells in other regions. Specifically, we require that the selected gene be both significant in one-versus-one tests (comparing it to one domain at a time, and pass significance threshold $P < 0.05$ in at least seven of eight such tests, Welch's *t* test) and significant in one-versus-rest test ($P < 1 \times 10^{-6}$, Welch's *t* test). The use of *t* test is justified as the expression z-scores are approximately normally distributed (**Supplementary Fig. 26**). Non-parametric Mann–Whitney U tests yield similar signatures (**Supplementary Fig. 27**). Accordingly, we defined a metagene signature as the average expression level for this subset of upregulated genes. These domain-associated metagene signatures (**Fig. 2d**) transcend cell types (**Supplementary Figs. 6** and **7**). Furthermore, we restricted this comparison to each specific cell type, and obtained an additional list of genes that are differentially expressed between domains. An expanded domain-metagene signatures was then defined based on the merged gene subsets. For glutamatergic cells, the expanded metagene signatures are summarized in **Supplementary Table 5**.

Analysis of spatial structure in the single-cell RNAseq data. To systematically characterize the spatial structure in a scRNAseq data, we summarized the gene signature associated with each spatial domain as a metagene (as described in the previous section). For simplicity, the overall expression of an expanded domain-specific metagene signature in each cell was quantified as the mean z-scored expression of all constituent genes in the signature. A t-SNE analysis was performed on this matrix using the Rtsne package with parameters pca_scale = T, perplexity = 35. Cell subpopulations with similar metagene expression patterns were identified by *K*-means clustering analysis (*K* = 9). We next annotated each cluster as belonging to the expression of one metagene. By comparing the binarized metagene expression population (**Fig. 4b**) and the *K*-means cluster annotations (**Fig. 4a**), all of the *K*-means clusters were assigned as uniquely associated with a single metagene.

For each subpopulation discovered from metagene clustering above, we found differentially expressed (DE) genes for the population (two-sample *t* test, unequal variance, $P < 0.05$). With the DE genes, we carried out Gene Ontology enrichment analysis (using hypergeometric test) for each of the nine subpopulations to construct a functional enrichment profile (hypergeometric test, top 500 DE genes analyzed per group, multiple hypothesis[51] corrected $P < 0.05$; **Fig. 4**). Here we used genes expressed in glutamatergic cells as the background gene-set when doing enrichment analysis.

Ref. 27 also provides layer information for a glutamatergic cell subset based on the layer from which the cells were manually dissected using different Cre-lines. To test whether the extracted subpopulation based on metagenes is enriched for a certain manually dissected layer of cells, we also performed hypergeometric test corrected for multiple hypothesis comparing manual annotations of cells to our HMRF-domain-based annotations.

Visualization of spatial domain and cell-type-specific variations. We created box plots to visualize the range of expression values for cells in different domains and for different cell types. In addition, to evaluate cell type transcending effect of domain signature genes, for each gene, we grouped cells by (cell type, spatial domain) pair, and plotted the expression distribution across groups ordered by spatial domains. Groups with less than four cells were removed, as these skewed the comparison.

Morphological analysis. We loaded the cell segmentations as regions of interest files (ROI) in ImageJ[52], then used the Measure tool available in ImageJ to quantitatively measure over 15 morphological features for individual cells. We compared the distributions across different cell types by using the Kolmogorov–Smirnov test. Statistical significance is judged by both 1) significance in at least seven of eight one-versus-one tests ($P < 0.05$ per test), and 2) significance in one-versus-rest test ($P < 0.0001$).

doi:10.1038/nbt.4260

Code availability. Code has been deposited at https://bitbucket.org/qzhudfci/smfishhmrf-py.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability. Expression data, spatial coordinates, SVM predictions, HMRF domains, expression box plots categorized by domains and cell types, and interactive visualizations are available at http://spatial.rc.fas.harvard.edu. The scRNA-seq dataset referenced in this study is GSE71585.

39. Caicedo, J.C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).
40. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
41. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
42. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
43. Li, S.Z. Modeling image analysis problems using Markov random fields. in *Handbook of Statistics* Vol. **20**, 1–43 (Elsevier Science, 2003).
44. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
45. Obayashi, T. & Kinoshita, K. COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res.* **39**, D1016–D1022 (2011).
46. Moffat, A. & Zobel, J. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* **27**, 1–27 (2008).
47. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B.* **63**, 411–423 (2001).
48. Dempster, A.P., Lamb, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38 (1977).
49. Feng, D., Tierney, L. & Magnotta, V. MRI tissue classification using high-resolution Bayesian hidden Markov normal mixture models. *J. Am. Stat. Assoc.* **107**, 102–119 (2012).
50. Brélaz, D. New methods to color the vertices of a graph. *Commun. ACM* **22**, 251–256 (1979).
51. Storey, J.D. & Tibshirani, R. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
52. Schneider, C.A., Rasband, W.S. & Eliceiri, K.W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

nature research

Corresponding author(s): Guo-Cheng Yuan

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a | Confirmed

☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☒ ☐ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	None.
Data analysis	We use custom algorithms to analyze the data. The software package developed in this study has been deposited at https://bitbucket.org/qzhudfci/smfishhmrf-py. Other packages are described in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Expression data, spatial coordinates, SVM predictions, HMRF domains, and expression box-plots categorized by domains and cell types, and interactive visualizations are available at https://spatial.rc.fas.harvard.edu. The scRNAseq dataset referenced in this study is GSE71585.

Nature Biotechnology: doi:10.1038/nbt.4260

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	One mouse brain was assayed by seqFISH due to experimental cost. No sample size was calculated.
Data exclusions	No data were excluded from analysis.
Replication	The reproducibility of seqFISH assays was established elsewhere in previous publications (Lubeck et al. 2014; Shah et al. 2016a, Shah et al 2016b). Here only one mouse brain was assayed.
Randomization	No randomizaton was considered.
Blinding	No blinding was considered.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	C57BL/6 with Ai6 Cre-reporter (uncrossed) (Jackson Laboratories, SN: 007906) female mice aged 50–80 days were anesthetized with isoflurane according to institute protocols (protocol #1701-14) (Madisen et al., 2012).
Wild animals	None.
Field-collected samples	None.