METHOD

Open Access

SMORE: spatial motifs reveal patterns in cellular architecture of complex tissues

Zainalabedin Samadi[1], Kai Hao[1] and Amjad Askary[1*]

*Correspondence:
amjada@ucla.edu

[1] Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles 90095, CA, USA

Abstract

Deciphering the link between tissue architecture and function requires methods to identify and interpret patterns in spatial arrangement of cells. We present SMORE, an approach to detect patterns in sequential arrangements of cells and examine their associated gene expression specializations. Applied to retina, brain, and embryonic tissue maps, SMORE identifies novel spatial motifs, including one that offers a new mechanism of action for type 1b bipolar cells. Structural signatures detected by SMORE also form a basis for classifying tissues. Together, our method provides a new framework for uncovering spatial complexity in tissue organization and offers novel insights into tissue function.

Keywords: Spatial transcriptomics, Neighborhood graphs, Uniform path sampling, Network motifs

Background

A central theme in biology is the idea that function is shaped by structure. Biological tissues, for example, often comprise stereotyped organizations of specific cell types that together enable proper function of the tissue. Formation of these structures during development is orchestrated by intrinsic gene regulatory networks and extrinsic cell-cell interactions. Therefore, analysis of cellular architecture of tissues can provide insight into both developmental processes that generate them and mechanisms that govern their function.

Recent technological developments in spatial omics have enabled researchers to map the position of cell types in complex tissues [1, 2]. Developing methods to analyze and interpret the resulting cellular maps is an active area of research [3–5]. Reported developments can roughly be classified into three groups. First, methods that examine spatial distribution of one cell type label relative to itself. This includes spatial autocorrelation [6] and Ripley's spatial statistics [7] that determine whether cells with a given label are clustered, dispersed, or randomly distributed in space. Second, methods that quantify association between pairs of cell type labels based on proximity. Analysis of this kind,

which is implemented in popular tools such as [5, 8] and also utilized in several other studies [9–13], can reveal enrichments or depletions in cell-cell interactions by comparing frequency of pairwise interactions in the samples with a random configuration. Third, methods that identify distinct cellular neighborhoods or microenvironments [14–21]. These methods typically cluster cells based on an embedding that represents the types and abundances of their neighboring cells. Variations of this approach have been used to study immune tumor microenvironment in colorectal cancer [14] and reorganization of local tissue architecture in response to acute kidney injury [21]. Methods have also been developed to analyze higher-order assembly of cellular neighborhoods, such as constructing "Tissue Schematics" [15] or to employ concepts from natural language processing, like "bag-of-words" idea in Spatial-LDA, to identify distinct microenvironment "topics" [16].

Despite the remarkable diversity of existing methods for spatial analysis, to our knowledge, none of them capture patterns in sequential arrangement of cells, as they focus on composition of cell types in each region regardless of their order. Spatial ordering of cell types within tissue is crucial for understanding organizational principles. Stereotypical sequential arrangements appear in many tissues, for example in the intestinal crypts along the crypt-villus axis, in airway epithelial cells of bronchioles, and in layers of skin epidermis. But their intricacy and significance is perhaps most evident in the central nervous system, where sequential arrangement of cell types enables precise signal processing and thus is directly related to tissue function. While some simple patterns in the spatial arrangement of cells are readily recognizable, the vast diversity of cell types revealed by spatial omics, combined with the complexity of biological tissues, necessitates a systematic statistical approach to uncover many of the underlying spatial patterns.

Here, we introduce a strategy for identifying "Spatial Motifs" that reveal statistically overrepresented spatial arrangements in complex tissues. Spatial omics data are often modeled as neighborhood graphs. Our approach focuses on paths in these graphs to directly capture ordered arrangements that may be overlooked by methods focused solely on regional composition. To extend the concept of motifs to spatial maps of cell types, we first developed an algorithm for enumeration and uniform sampling of paths in neighborhood graphs. Each path consists of a sequence of nodes, labeled by the cell types they represent. Paths along which the physical distance monotonically increases capture arrangements of cell types in the sample. Identifying overrepresented patterns in linear sequences, such as DNA, has long been a cornerstone of bioinformatics, with methods extensively refined and optimized over the years. To identify recurring patterns of cell types in sampled paths, we adapted the STREME algorithm, originally developed for motif discovery in nucleic acid sequences, enabling us to build on a history of highly optimized techniques (Fig. 1a). Our new algorithm, called Spatial MOtif REcognition (SMORE), introduces crucial modifications to accommodate input from spatial graphs rather than one-dimensional sequences. It also integrates motif discovery with differential gene expression analysis to compare cells within spatial motifs to those of the same type located elsewhere in the tissue.

We tested the sensitivity, specificity, and accuracy of SMORE by recovering motifs that were embedded at specified frequencies within synthetic graphs. We then
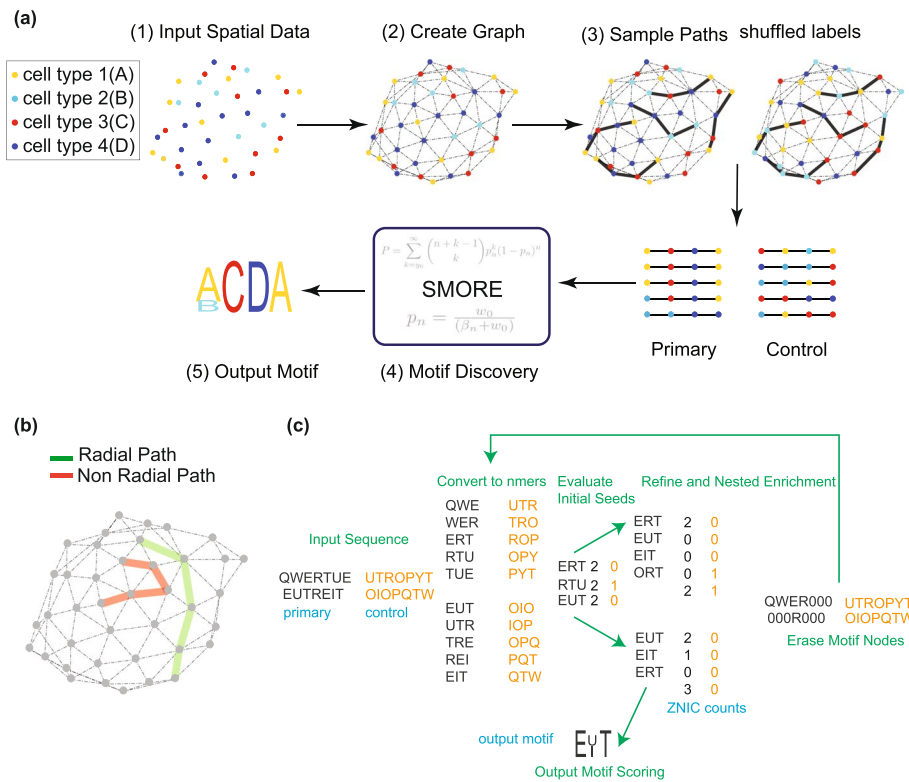
Fig. 1 Overview of the spatial motif discovery algorithm. **a** Schematic of the procedure for finding motifs on a set of spatially distributed nodes, each labeled by a color and a letter for illustration purposes. (1) A set of spatially distributed nodes in 2D space. (2) Delaunay triangularization is used in this case to generate the graph. (3) URPEN is used to sample a set of length-4 radial paths. Control data is created by shuffling the observed graph node cell types. (4) SMORE is used to extract most significant spatial motifs. (5) PWM logos can be used to represent the output motifs. **b** An example of a radial and a non-radial path. Radial path sampling ensures nodes that are farther from each other in the sample are also farther from each other in physical space. **c** A simplified example for the SMORE pipeline. Input samples are converted to nmers from length 3 up to length of the motif and p values of the seeds evaluated based on negative binomial test with Bernoulli probability computed based on total number of distinct seeds in primary and control graph. nEval = 3 most significant seeds are input to refine and enrichment where ZNIC p values are used to refine and enrich candidate seeds. Hold-out scoring is performed to compute the p value for the output motif. Nodes of the seeds involved in the motif are erased from the graph and the process is repeated again. More detailed description of each block is provided in the "Methods" section. Letters and sequences are chosen for demonstration purposes and do not correspond to specific cell types

analyzed published datasets of mouse retinal bipolar cells, hypothalamic preoptic region, and embryonic tissue to identify a variety of spatial motifs. Our results revealed that gene expression of cells in some spatial arrangements can differ significantly from other cells of the same type, providing clues to the functional significance of the spatial motifs. We also demonstrate SMORE's remarkable versatility and scalability by applying it across diverse spatial transcriptomics datasets, spanning both 2D sections and 3D volumes, imaging- and sequencing-based techniques, and a whole mouse brain dataset with nearly 4 million cells. Together, this work presents a novel and broadly applicable approach for identifying patterns in spatial data that go beyond pairwise associations and regional composition.

Design of the spatial motif discovery algorithm

Our algorithm for finding spatial motifs in neighborhood graphs consists of two main components: first, a method to uniformly sample paths from the graph, and second, a procedure to find motifs in the obtained samples. Each path provides a sequence of cell type labels that occur near each other in space. Generating sequence samples from the neighborhood graphs reduces identification of spatial patterns into finding overrepresented sequences of labels. Despite some important differences, this task is similar to identifying motifs in nucleic acid or protein sequences. Therefore, we generalized existing methods of motif discovery in genomic sequences for our application on graphs.

The sampling algorithm takes a graph G and returns an unbiased sample of all paths inside the graph. In a graph, a path is a walk that does not intersect itself. Our selected paths are also constrained to be "radial." Radial condition in a spatially embedded network is defined as the requirement that physical distance along a path monotonically increases along the sequence of edges in the path. Radial condition ensures that the sequence of labels in a sampled path corresponds to a spatial arrangement of cell types in space, so that labels that are farther from each other in the path are also farther from each other in physical distance (Fig. 1b). Therefore, the radial requirement simplifies interpretation of the output motifs.

After sampling, the motif discovery algorithm identifies sequences of a given length that are statistically overrepresented in an iterative process. In each step, a significant recurring pattern of cell types is identified and is subsequently refined by considering sequences similar to the initial pattern or seed (Fig. 1c). The algorithms for path sampling and motif discovery are described below. For more details, refer to the extended "Methods" section. The time required for analysis of the experimental datasets in this study is summarized in Additional file 1: Table S1. The code for sampling and motif discovery algorithms is available at: SMORE: Spatial Motif Recognition.

Sampling the graph to generate paths

The first step in our approach involves uniform sampling of neighborhood graphs. We have developed an algorithm for Uniform Random Path Enumeration (URPEN) based on the Rand-ESU algorithm [22]. The Rand-ESU method involves enumerating all potential subgraphs within a given graph, incorporating a probability element to uniformly sample a subset of these subgraphs. In our modification, we have adapted this method to exclusively sample paths as opposed to subgraphs. Paths differ from subgraphs in that they cannot intersect themselves, and each node, excluding the initial and terminal nodes, is only linked to its preceding and succeeding nodes in the path sequence. This contrasts with subgraphs where nodes can be connected to an arbitrary number of neighboring nodes. This distinction is crucial when selecting the next neighbor to expand the growing sample.

Enumerating all length-k paths in a graph

The Path Enumeration algorithm (PEN) (Algorithm 1) enumerates all paths of length k within a graph. The algorithm begins with a vertex v from the input graph and adds only those vertices to the set that are neighboring the newly added vertex w but are not

already in V_{path}. To prevent enumeration of both the path and its reverse, for undirected graphs, the index of the last vertex in the enumerated path must be greater than that of v.

Uniform path sampling

Similar to ESU-tree, the PEN algorithm's structure can be visualized as a tree structure. The tree structure for an example graph is demonstrated in (Fig. 2a). This tree has 18 leaves which correspond to the 18 size-3 paths of the graph. We can use this tree to sample paths uniformly without bias. The PEN algorithm systematically traverses its associated PEN-tree. In situations where a full traversal is impractical, we can perform
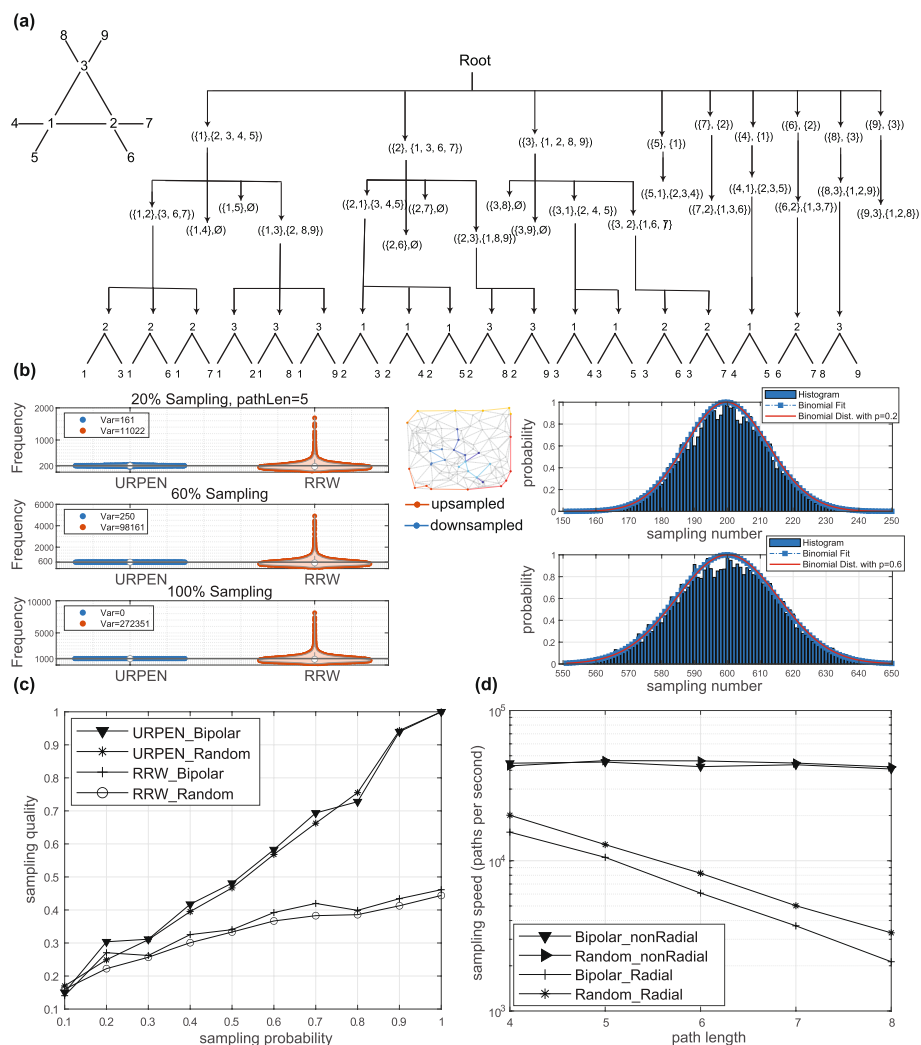

Fig. 2 Unbiased sampling of paths from neighborhood graphs. **a** An example of the PEN sampling tree. **b** URPEN returns each path at a frequency corresponding with the sampling level, whereas Radial Random Walk (RRW) results in biased sampling of the graph. Examples of RRW upsampled and downsampled paths are shown in the top right corner of the 20% sampling panel. The sampling probability in URPEN is set to $p = (1, 1, \ldots, 1, 0.2)$ for 20% sampling, and $p = (1, 1, \ldots, 1, 0.6)$ for 60% sampling. This test was performed 1000 times. **c** Sampling quality with URPEN and RRW on the bipolar cell type graph and a random graph with 12000 nodes. **d** Sampling speed for non-radial and radial sampling on the bipolar and random networks

a partial exploration of the PEN-tree such that each leaf is reached with equal probability. To achieve this, a probability is introduced for each depth $1 \leq d \leq k$ in the path (or each depth in the PEN-tree), and the subsequent node rooted at a node at depth d is traversed with probability p_d. This is implemented by calling the ExtendPath function at lines 3 and E6 of the PEN algorithm (Algorithm 1) with probability p_d. This new algorithm is called Uniform Random Path Enumeration, URPEN. It can be observed that URPEN visits each path with equal probability of $p = \prod_{d=1}^{k} p_d$. The method is tested in the "URPEN enables efficient and unbiased sampling of neighborhood graphs" section on a random graph to validate its accuracy numerically.

Spatial MOtif REcognition (SMORE)

Applied to a neighborhood graph, URPEN returns sequences of cell types that are observed near each other. Similar to DNA sequences, we can search through these cell type sequences for motifs using unsupervised learning techniques. The SMORE method is developed to detect motifs within the sampled sequences. Our approach uses as its basis the recently developed STREME method [23] which the author has demonstrated to be more accurate, sensitive, and comprehensive than several widely used motif discovery algorithms. STREME is developed to find motifs within sequence-like samples; SMORE on the other hand finds motifs in a network-based dataset. The algorithm follows a series of steps to accomplish its objectives.

1. Construct the graph from spatial data: To construct a graph from the spatial coordinates of cell types, Delaunay triangularization is employed, forming a graph with nodes as cells, labeled with their respective cell types. Given that each dataset may comprise multiple tissue sections or animal IDs, separate graphs are generated for distinct sections and IDs. For generating control data, each section or animal ID is shuffled independently. Additionally, besides Delaunay triangularization, the method offers options to construct the graph using arbitrary K nearest neighbors or epsilon graph approaches.

2. Sampling the graph and generating control data: SMORE uses URPEN for uniform sampling of the input graph. Control samples are generated using one of two methods: shuffle or kernel. The shuffle method produces control data by shuffling node cell type labels within samples (e.g., tissue sections and animal IDs). On the other hand, the "kernel" method establishes a kernel around each cell and swaps the cell's label with that of a cell within its kernel (Additional file 1: Fig. S1). In the experiments detailed in this paper, kernels with a radius of K neighbors are used, where K is specified for each experiment. $K = 1$ means only first neighbors within the graph are considered for shuffling. The degree of randomness in the control data is controlled by adjusting the number of neighbors considered for shuffling.

   There are certain scenarios where specific cell types associated motifs are readily apparent. In such cases, these cells can be fixed between primary and control data, meaning that their cell labels are not shuffled.

   We generate nTrain control data and seed numbers for control data are the total number of that specific seed within these nTrain samples. NScore independent con-

trol data is generated for output motif scoring, with the same settings as the training data.

3. Convert to n-mers and count seeds: Input samples to the algorithm can be of an arbitrary length. These samples are converted to W-mers, where W is the desired motif length. These W-mers are input to the count seed modules. Three strategies have been proposed for counting motif instances: counting all occurrences, counting those with no shared edges, and counting those with no shared nodes [24]. We adopted the third approach, which assumes that motifs share no common nodes, to prevent overrepresentation due to overlapping occurrences. This approach, referred to here as the Zero Node in Common (ZNIC) model ensures that each counted motif instance is structurally independent. Motif counts under ZNIC model are referred to as ZNIC counts. By using the ZNIC model, we avoid inflated frequency counts, allowing for a more accurate capture of the network's true structural patterns. ZNIC counts of unique seeds, along with their associated nodes, are then passed to the next module for evaluating initial seeds.

4. Initial seed evaluation: The significance of each initial seed is obtained by the negative binomial test (see the "Methods" section for details). The justification for this specific test is argued in the extended "Methods" section. The first nEval seeds with this criterion are passed to the next stage of refinement and enrichment. The default value for nEval is 25.

5. Refinement and nested seed enrichment: The refinement and seed enrichment both use the same process of enrichment, except that refinement is only one iteration, and seed enrichment is nEnrich iterations, with the default value of nEnrich = 20. Enrichment groups similar path samples together and compares ZNIC counts of the grouped sequences with the control data. nEval motifs from the initial evaluation step are first enriched for one iteration and top nRef (nRef = 4 as a default) motifs are further refined in seed enrichment block for nEnrich iterations or until p value does not improve.

   During each iteration, the Position Weight Matrix (PWM) is calculated from the sequences participating in the motif. Likelihood ratio scores are then computed for all samples, using this PWM matrix, and the samples are arranged in descending order based on their PWM scores. In the event of an equal PWM score, the samples are further sorted based on their p values obtained in the initial evaluation block. Subsequently, ZNIC counts for the ordered samples are determined, and the PWM score threshold that minimizes the p value is identified. This process is iterated if the p value obtained is more significant than the previous iteration.

6. Motif scoring: In most cases, tissue samples are different from each other and it is not optimal to take sections of the samples as hold-out for scoring. In our experiment, the same samples are used for finding motifs and scoring the output motif, with the scoring part iterated over NScore times with different shuffled networks to avoid false positives. Implementation results on synthetic data with nScore = 50 shows that the false positive rate is negligible. Furthermore, tests on real data with randomly shuffled cell type labels did not result in any significant output motif. The ZNIC counts for the seeds involved in the output motif are computed in the primary

and nScore randomly generated control data and the 95 percentile least significant *p* value is considered as the output *p* value of the scoring module.

7. Motif node erasing: The respective nodes for the seeds involved in the motif are erased (i.e., their cell type is set to 0) from the graph in primary and control data along with their reverses. The previous steps are then repeated until output motifs are not significant anymore, or a specified number of output motifs have been discovered.

Results

URPEN enables efficient and unbiased sampling of neighborhood graphs

The sequence of cell types along a path in the neighborhood graph captures their local spatial arrangement. A collection of these sequences can be used for identifying overrepresented patterns in the graph, only if it represents an unbiased sample of all possible paths. Furthermore, the sampling algorithm should be able to handle the large number of cells in typical spatial transcriptomics datasets.

To confirm that URPEN sampling is unbiased, we generated a graph by applying Delaunay triangularization on a spatially uniform distribution of 120 two-dimensional Cartesian points. We then used URPEN to sample radial paths of length 5 from this random graph at three sampling levels; 20%, 60%, and 100%. If sampling is unbiased, we expect each path to appear in the sample with a probability equal to the sampling level. We repeated these tests 1000 times. So, for the case of 20% sampling, we expect each path to appear on average 200 times in the output results. This number is 600 for 60% sampling and 1000 for complete 100% sampling. As a comparison, we also sampled the same graph with the commonly used random walk method. Random walk starts with a randomly chosen node and subsequent nodes are selected from the neighboring nodes with equal probability, until a path of the desired length is obtained.

While URPEN returned paths at the expected frequency, random walk sampling showed significant bias for certain paths (Fig. 2b). The distribution of the URPEN counts is also consistent with a set of identical independent binomial distributions with $p = p_d$, confirming that paths were sampled with equal probability (Fig. 2b). Similar results were obtained for other path lengths and for paths not constrained by the radial condition (Additional file 1: Fig. S2). The sampling bias of random walk can be mitigated by using unbiased estimators [25]. However, that incurs complexity and leads to a penalty in speed performance [22].

We also evaluated the sampling quality of URPEN compared to random walk (Fig. 2c). Sampling quality is defined as the percentage of path types for which the number of extracted paths has at most 20% error relative to the exact counts, similar to the measure used previously [22]. Path samples with at least 5 counts were considered for quality evaluation. Sampling quality was computed for two cases: A random graph with 12,000 nodes and 35,823 edges and a graph based on spatial distribution of bipolar cells in a section of mouse retina [26]. Sampling quality for the URPEN method increases with increasing sampling probability, reaching one at complete sampling. In contrast, sampling quality for the random walk peaks at around 0.5.

We then assessed how sampling speed scales with the path length (Fig. 2d). The two graphs of Fig. 2c was sampled by URPEN at 10% level. The speed appeared to be

independent of the path length for non-radial paths. However, when the radial constraint was applied, the speed decreased with path length, because the proportion of paths that meet the radial condition decreases with increasing path length. Consistent with this explanation, the speed of sampling radial paths is not independent of graph architecture, as it appears to be the case for the non-radial paths. Together, our results demonstrate that URPEN avoids the shortcomings of random walk sampling and offers a robust method for uniform path sampling.

SMORE accurately identifies overrepresented cell type arrangements in synthetic graphs

Evaluating the performance of SMORE requires datasets with known ground truth. Since the presence and frequency of overrepresented spatial patterns in existing experimental data is unknown, we created synthetic data by embedding patterns within random graphs at known frequencies. Each graph has 12,000 nodes, 35,823 edges, and 12 cell types. The embedding percentage indicates the proportion of nodes used for pattern embedding. For example, 2% embedding indicates that 2% of the nodes (i.e., 240 nodes) were used for embedding patterns. In the case of a length 4 motif, this results in 60 sequence patterns. Other nodes in the graph were labeled randomly. Embedded patterns included one variable position, which was filled with one of two cell types with equal probability. Other positions in the pattern were assigned one predefined cell type. Although more complex patterns can also be considered, we chose our patterns so that they are sufficiently complex while still providing insight into the algorithm's performance. The algorithm was run 100 times for each embedding frequency, generating 10 output motifs per run. The samples for length 4 motif included all possible radial paths inside the graph. For the length 5 motif, the graph was downsampled by URPEN and 65% of all radial paths were used.

The accuracy of a motif discovery algorithm describes its ability to recover accurate versions of overrepresented patterns [23]. We assessed the accuracy of SMORE by measuring the similarity of output motifs to embedded patterns (Fig. 3a). In each run, SMORE returns 10 motifs that are sorted based on their statistical significance. At embedding frequencies above 0.5%, the most significant output motif tends to be the one most highly correlated with the embedded pattern. At 1% and higher embedding frequency, the Pearson correlation coefficient between the extracted motifs and the embedded patterns was higher than 0.9 in all 100 tests. Based on these results, we expect SMORE to be able to accurately identify length 4 and 5 motifs even when they only occur at low frequencies in a sample.

To assess the specificity of SMORE, we examined the true positive rate (TPR) and false positive rate (FPR) against output p values (see the "Methods" section for details). Ideally, both TPR and FPR should be high at high p values and decrease as the p values decrease, with FPR decreasing at a faster rate to allow for correct motif identification. Accordingly, we observed that FPR curves reached zeros at around log p value of -10, while TPR performance gradually increases with the embedding frequency (Fig. 3b). The corresponding ROC curves (Fig. 3c) confirm close to perfect classification performance of SMORE at embedding frequencies higher than 1%.

To evaluate the sensitivity of SMORE, we defined success rate to be the proportion of output motifs that are statistically significant, at a given p value, and have a Pearson
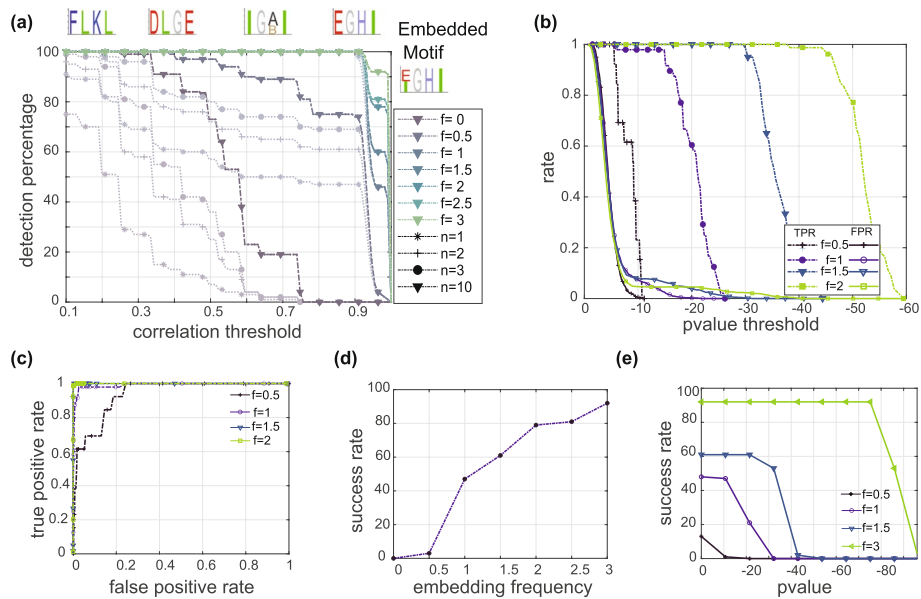
Fig. 3 Evaluation of SMORE's performance on synthetic data with known ground truth. **a** Accuracy of SMORE, defined as the best correlation coefficient between the embedded motif and extracted motifs for length 4 motifs with varying embedding frequency (*f*). This figure represents the average results from 100 algorithm runs. The highlighted curves indicate the best correlation among the $n = 10$ output results. Faded curves demonstrate the results when considering only $n = 1, 2$, or 3 of the best output motifs, instead of all 10. Examples of motifs with different levels of correlation are shown on the top. **b** Specificity results for length 4 motifs, measured by false positive rate and true positive rates against log *p* value threshold and (**c**) against each other. **d** Sensitivity of SMORE measured by successful motif recovery rate against embedding frequency and (**e**) enrichment log *p* value (see Additional file 1: Fig. S3 for length 5 results)

correlation coefficient of at least 0.95 with respect to the embedded pattern. With a log *p* value threshold of −10 (Fig. 3d), success rate for identifying length 5 motifs exceeds 90% at embedding frequencies above 1%. Success rates for length 4 motifs increase more gradually with embedding frequency, exceeding 90% only at 3% embedding. At each embedding frequency, success rate appeared to drop sharply beyond a specific *p* value threshold (Fig. 3e). As expected, this threshold decreases with increasing embedding frequency.

SMORE reveals spatial motifs in the distribution of mouse retinal bipolar cells

Retina contains a rich diversity of neuronal cell types, organized into three layers of cell bodies. These cell types have different features and frequencies and are tiled across the retina in a stereotypical manner that supports the overall function of the tissue. Cell type diversity and individual variability make it difficult to identify recurring patterns in the cellular architecture of the retina. Spatial motif analysis can reveal higher order associations between retinal neurons and provide insight into their development and wiring.

We applied SMORE on a dataset of the mouse bipolar interneuron subtypes containing more than 30,000 cells [27]. These subtypes were differentiated using co-detection of 16 gene markers by SABER-FISH, allowing the classification of all 15 bipolar subtypes. Bipolar interneurons bridge all visual circuits, establishing the link between sensory rod and cone photoreceptors and the output neurons. Bipolar cells

also do not migrate from their birthplace, providing a spatial map between their final location and the location of their progenitors [28].

Rod bipolar cells (RBCs) constitute the majority of retinal bipolar cells in mice and their cell bodies are mainly organized together, further out in the inner nuclear layer (INL) compared to cone bipolar cells (CBCs) (Fig. 4a). SMORE evaluates the statistical significance of cell type arrangements in the experimental data against control data, which are generated by shuffling cell type labels. When shuffling is done for all cells across the whole graph, relatively obvious structures, like separation of RBCs, are identified as highly significant motifs (Fig. 4b, motif #2). To reveal other motifs involving RBCs, besides this trivial case, we can fix the position of RBCs in the control data (Fig. 4a). This would eliminate any motif whose significance stems from RBCs and elucidate the relationship between RBCs and other cell types (Fig. 4c, motifs #1 and 3).

Our analysis revealed several highly significant spatial motifs among retinal bipolar cells. These motifs can be investigated in the context of retina development, anatomy, and physiology. For example, when RBC positions are fixed, the most significant motif involves a type 2 OFF CBC followed by two RBCs and a type 6 ON CBC (COOK motif; Fig. 4c). Overrepresentation of this cellular arrangement can be understood in the context of a primary rod pathway that enables scotopic vision [29]. Within this pathway, the signal originating from a single rod cell is primarily directed to a select few AII amacrine cells through two RBCs [30]. The AII amacrine cells establish connections with nearly all bipolar cell types to gather scotopic signals originating from RBCs (denoted as O in our motif notation). These signals are then distributed to both ON and OFF CBCs through gap junctions and inhibitory synapses, respectively (Fig. 4d). However, the number of connections varies depending on the bipolar cell types [31]. Type 2 (C) cells account for 69% of the total number of OFF bipolar chemical synaptic contacts with AII amacrine cells, while type 6 (K) cells contribute 46% of the total area of ON bipolar gap junctions with AII amacrine cells [31]. Both type 2 and type 6 cells not only have the highest access to AII amacrine cell signals but also share these signals with other types of bipolar cells through interconnected gap junctions in the network. These findings support the central role of type 2 and type 6 cells in conveying the most sensitive scotopic signals to the postsynaptic ganglion cells. Furthermore, AII amacrine cells are characterized by their narrow-field dendrites. Typically, a bipolar cell receives more inputs from AII amacrine cells that are in its close proximity [31]. This suggests that bipolar cell types involved in scotopic vision should be spatially close to each other. Given these considerations, we hypothesize that the COOK motif is associated with the primary rod pathway for scotopic vision in mice.

SMORE is specifically designed to identify overrepresented sequential arrangements of cell types. Therefore, its objective and output differs from previous methods that search for spatial neighborhoods or microenvironments based on local cell type composition, regardless of order of cells. To clarify this distinction, a side-by-side comparison between SMORE and two such methods, HistoCAT [8] and ImaCytE [19], is included in the Supplementary Information (Additional file 1: Fig. S4).
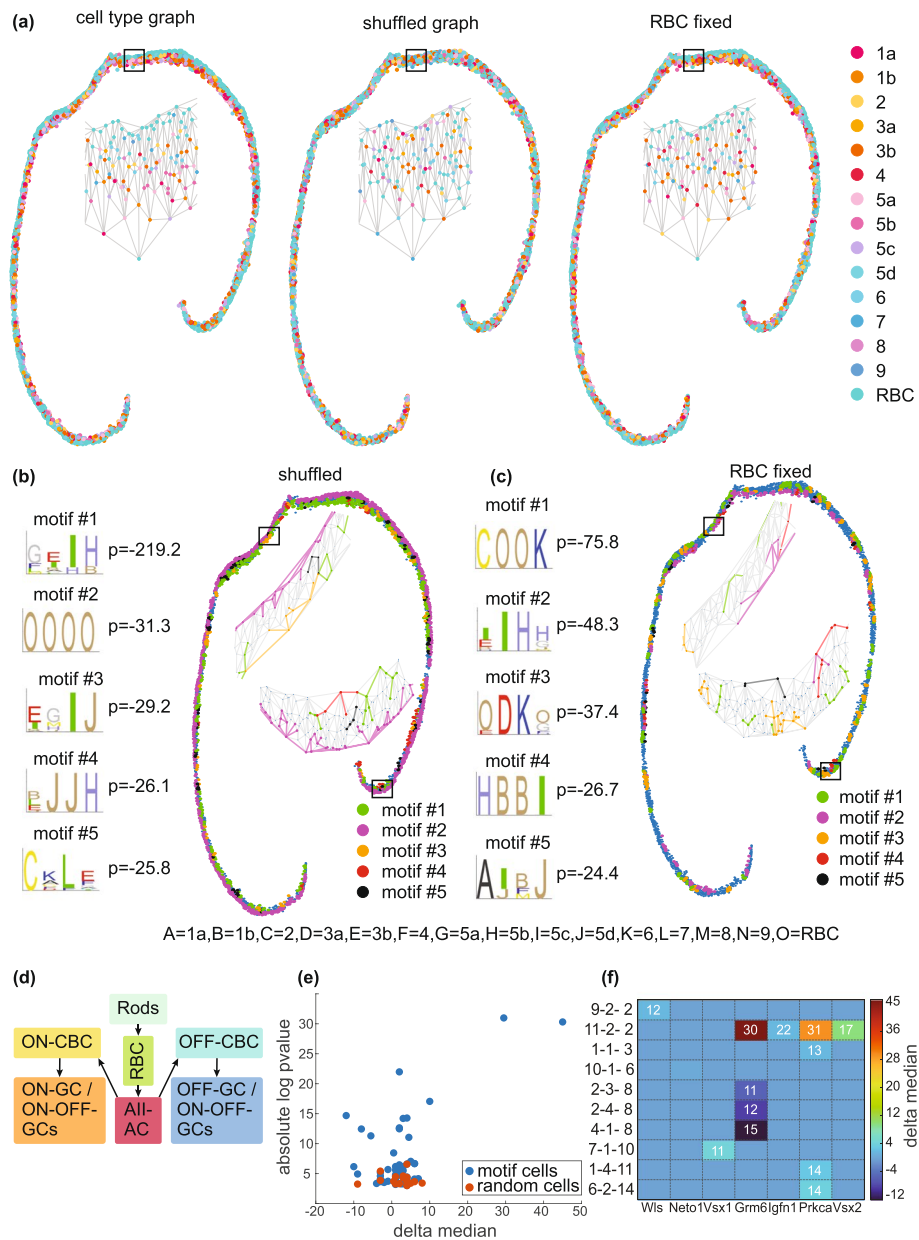
Fig. 4 Spatial motif analysis of mouse retinal bipolar cells. **a** A retinal section with classified bipolar subtypes (left) and examples of control data generated from this section using different randomization methods, global shuffling (middle) and shuffling with fixed RBCs (right). **b** The top five output motifs, identified using the global shuffling method, are displayed in order from top to bottom with their respective log *p*-values. Their positions within a section of the retina are shown by highlighting the nodes associated with each motif, using the color code in the bottom right corner. Two example regions, marked by rectangles, are enlarged for a closer view. **c** Same as **b**, but for motifs obtained when RBCs are fixed. Annotations for the cell types involved in the motifs are listed at the bottom. **d** Schematic for primary pathway for rod-driven signals involving rods, rod bipolar cells, AII amacrine cells, OFF or ON (cone) bipolar cells, and OFF or ON ganglion cells. **e** Absolute log *p* value versus difference in gene expression medians (delta median) for cells in a spatial motif versus cells of the same type that are not in a motif arrangement. The results for a random selection of cells are shown in red. **f** Selected cases and genes with absolute log *p* value greater than 7. The heatmap is colored with delta median values. The motif cases, specified by motif number-position-cell type, are sorted by cell type on the vertical axis. Values in each cell show absolute log *p* value for comparison between cells within a given motif and overall cells of the same type

Cells within spatial motifs exhibit gene expression differences compared to other cells of the same type

Spatial context of the cells often influences their function. With recent advances in spatial gene expression profiling, there has been an increased interest in systematically characterizing spatial variability of gene expression [32–38]. Spatially variable genes may explain functional distinctions between cells in different regions or demarcate spatial domains [32–34, 37]. If spatial motifs represent functional units made of various cell types that together play a distinct role, we can expect cases with distinct gene expression signatures. Specifically, cells within some spatial motifs may exhibit unique gene expression profiles compared to cells of the same type that lie outside these specific spatial arrangements.

We assessed differential gene expression between cells of each type that are involved in a spatial motif and the ones that are not. For each gene, the median expression value among the cells of a specific type that are involved in the motif was subtracted from the median of all the cells of the same type. The p values for the observed delta medians were obtained through theoretical computation by analyzing the distribution of delta median values for a random subset of cells (see the "Methods" section for details). We performed this analysis for motifs obtained by shuffling with fixed RBC cells for the 16 genes profiled in the retinal bipolar dataset [27]. Several cases of highly significant differential gene expression were observed in the motif cells (Fig. 4e). In contrast, control samples where a random subset of cells were selected, with the same size and type of the corresponding motif case, showed much higher p values. This observation is consistent with functional specialization of cells in spatial motifs.

The heatmap in Fig. 4f illustrates the absolute log p values across all cases for 20 output motifs. Each motif can consist of multiple cell types in different positions. For example, the second motif in the fixed shuffling case of Fig. 4c comprises 10 cells, 4 cells in each of the positions 1 and 4, and one cell in positions 2 and 3. Here, we consider each cell type in each position of each motif as a separate case. Genes with absolute log p values greater than 7 are highlighted in Fig. 4f.

Among genes that were significantly upregulated or downregulated in the motifs, Grm6 stands out because it shows the most extreme differential expression in both directions. Grm6 is upregulated in type 1b OFF bipolar cells (B) in OBBO motif, where O represents an RBC, (delta median $= 45, -\log(p\text{value}) = 30.31$) and downregulated in type 5b ON bipolar cells (H) in HBBI motif (Delta median $= -12, -\log(p\text{value}) = 14.67$). Grm6 encodes the metabotropic glutamate receptor 6 (mGluR6) which is localized to the dendritic tips of ON bipolar cells [39]. It plays a crucial role in triggering depolarization of ON bipolar cells in response to light-induced hyperpolarization of photoreceptors [40]. Mutations in Grm6 gene in humans lead to autosomal recessive congenital stationary night blindness (arCSNB) [41].

We observed that enrichment of Grm6 in type 1b cells in motif 11 (OBBO) can be explained by their position along the radial axis of the retina. Grm6 is expressed at higher levels in type 1b cells whose cell body is closer to the photoreceptor level (Fig. 5a). Since RBCs (O) are concentrated in this outer region, type 1b cells in OBBO motif also tend to be in radial positions where Grm6 expression is higher (Fig. 5a). In contrast, downregulation of Grm6 in type 5b cells (H) of motif 4 (HBBI) seems to be explained by
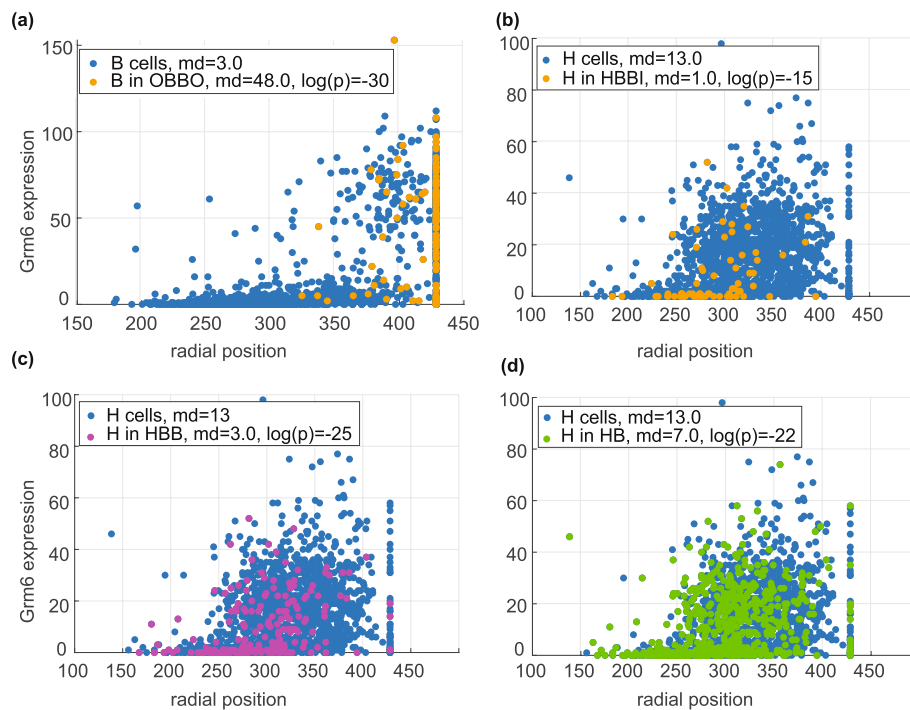
Fig. 5 Grm6 shows motif specific expression patterns. Expression of Grm6 versus radial position of cell in the mouse retina for (**a**) type 1b and (**b**–**d**) 5b cells. Cells within specific spatial arrangements are highlighted in each panel. The radial position was computed by considering the outer extreme points as the maximum radius points and computing the radial position for other nodes relative to the nearest outer point

their proximity to type 1b (B) cells rather than their radial position (Fig. 5b–d). Type 1b cells lack dendrites connecting them to photoreceptors [42, 43]. Therefore, the mechanism of their function is not well understood [42]. Our observation suggests that type 1b cells may influence signal processing in the retina by altering expression of key postsynaptic receptors, like Grm6, in nearby ON bipolar cells.

SMORE reveals overrepresented cell type arrangements in the preoptic area of mouse hypothalamus

To explore the utility of SMORE at identifying non-trivial recurring patterns in a significantly more complex sample, we applied it to a spatial transcriptomics dataset of the mouse hypothalamic preoptic region [44]. This dataset profiles about 1 million cells and has identified about 70 neuronal populations with distinct signatures and spatial organizations.

  As we showed before, the approach used to generate control data affects the output motifs. This can be used to tune the algorithm to different anatomical features. Here in addition to fixing specific cell types, we introduce local kernel shuffling (Fig. 6a–c). Shuffling cell labels across the entire sample can result in the emergence of relatively straightforward structural motifs, such as regional boundaries. On the other hand, local shuffling maintains cell type frequencies within cellular neighborhoods. Therefore, if a sample is compartmentalized to regions with distinct cellular compositions, local shuffling is more likely to identify patterns that are overrepresented within each compartment. Control
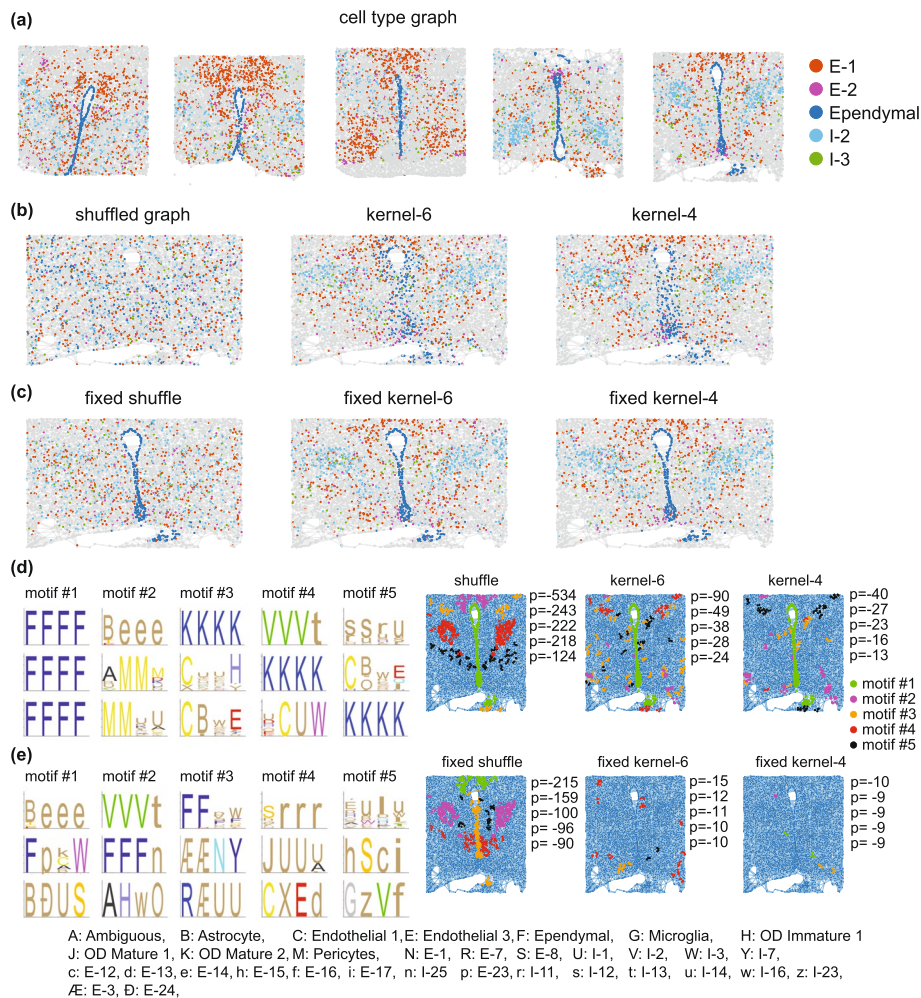
Fig. 6 Spatial motif analysis of mouse hypothalamic preoptic region. **a** The graph of the preoptic region of the hypothalamus from five adult male mouse brains at Bregma − 0.29. **b** Examples of control data generated from the 5th tissue using different randomization methods: global shuffling, kernel shuffling with 6 neighborhood depth, and kernel shuffling with 4 neighborhood depth. **c** Control data obtained by applying the same randomization methods as (**b**) but with the positions of non-neuronal cell types fixed. **d** The top five motifs, identified using global shuffling (top), kernel shuffling with depth 6 (middle), and kernel shuffling with depth 4 (bottom). The position of motifs within a tissue section is shown to the right, along with their log *p*-value, ordered from top to bottom. To show the position of the motifs, nodes associated with each motif are highlighted using the color code in the bottom right corner. **e** Same as **d**, but for motifs obtained when non-neuronal cell types are fixed. Annotations for the cell types involved in the motifs are listed at the bottom

data generated by local shuffling maintains a higher degree of similarity with respect to the original experimental data. Therefore, the motifs obtained with global shuffling tend to have more significant *p* values compared to the output motifs of local shuffling. In both shuffling methods, if certain cell types form obvious structures (e.g., Fig. 6a Ependymal cells (blue)), their positions can be fixed in the control data.

We applied SMORE to cellular maps of the hypothalamus preoptic region from five adult male mouse brains at Bregma − 0.29 to identify length 4 spatial motifs. The sections used in our analysis comprise 28,866 cells. We tried both global shuffling and local shuffling with kernel sizes of 4 and 6 (Fig. 6b). In another set of experiments, we also

fixed the position of non-neuronal cell types between the primary and control data (Fig. 6c). When non-neuronal cell types were not fixed, the most significant motif consists of a group of four interconnected ependymal cells (Fig. 6d). This pattern is immediately visible in the graphs because ependymal cells form a layer that lines the ventricles. As expected, fixing the position of non-neuronal cells removes this motif as well as the motif made entirely of mature oligodendrocytes (Fig. 6e). Instead, other motifs emerge, some of which are combinations of ependymal and neuronal cells (e.g., motif 3 of fixed global shuffling).

The first five motifs generated through the global shuffling method correspond to discernible patterns within the fifth tissue section shown in Fig. 6a. In addition to the first motif of ependymal cells, motif 2 comprises astrocytes, E-9, and E-14 cell types, which are enriched in the PVA nuclei of the hypothalamus [44]. E-14 and E-9 cells also show similar gene expression patterns [44]. Motif 3 represents a pattern of mature oligodendrocytes known to be enriched in the anterior commissure and the fornix. Motif 4 is a pattern of I-2 and I-13 cells which are indicated to be enriched in BNST-p and StHy nuclei. I-2 and I-13 are both aromatase-enriched clusters and express both androgen receptor (Ar) and estrogen receptor alpha (Esr1) [44]. Motif 5 is primarily composed of cell types I-11, I-12, and I-14, collectively enriched in the MPN and StHy nuclei.

Motifs obtained from kernel shuffling methods capture less obvious patterns. For instance, motif 5 in the case of "fixed kernel 6," represented as hSci sequence which is equivalent to a radial path of 4 excitatory neuronal cell types, E-15, E-8, E-12, and E-17, occurs 4 times in the primary data (2 occurrences in each of animal IDs, 10, and 11) and 3 times in the total of 50 generated control data. This motif consistently appears in kernel shuffled tests, with different ranks. Interestingly, there is a similar motif, BSic, with B representing astrocytes, that appears as motif 11 of the not fixed kernel 4 (Additional file 1: Fig. S5a) experiment and appears at the opposite side of the brains of the same Animal IDs.

The functional explanation of identified spatial motifs is not the focus of this study. But it is reasonable to expect that such patterns hint to either functional relationships between the cell types involved or specific developmental programs that generate them. Therefore, they can help generate hypotheses for future studies. For instance, motif 5 in the non-fixed global shuffling experiment primarily consists of cell types I-11, I-12, and I-14. In the case of the fixed shuffle, motif 4 is predominantly composed of cell types E-8, E-15, I-34, and I-15 in the first position, while I-11 occupies the remaining positions. I-15, I-2, I-11, I-14, I-33, E-8, and E-15 cells display sexually dimorphic cFos enrichment in male mating [44]. Interestingly, a similar motif exists in female sections (Additional file 1: Fig. S5b), where I-15 replaces I-11. The Esr1-enriched cluster I-15 exhibits significant enrichment in female animals and is preferentially activated in females, with lesser activation in males after mating [44].

We also performed gene expression analysis for the motifs obtained by global shuffling with fixed non-neuronal cells. Many cases of highly significant differential gene expression were observed in the motif cells (Fig. 7a). The heatmap in Fig. 7b illustrates the absolute log *p* values across all cases for 40 output motifs. The majority of statistically significant cases for genes imaged using combinatorial smFISH measurements are concentrated in the first 10 output motifs. In contrast, genes measured through sequential
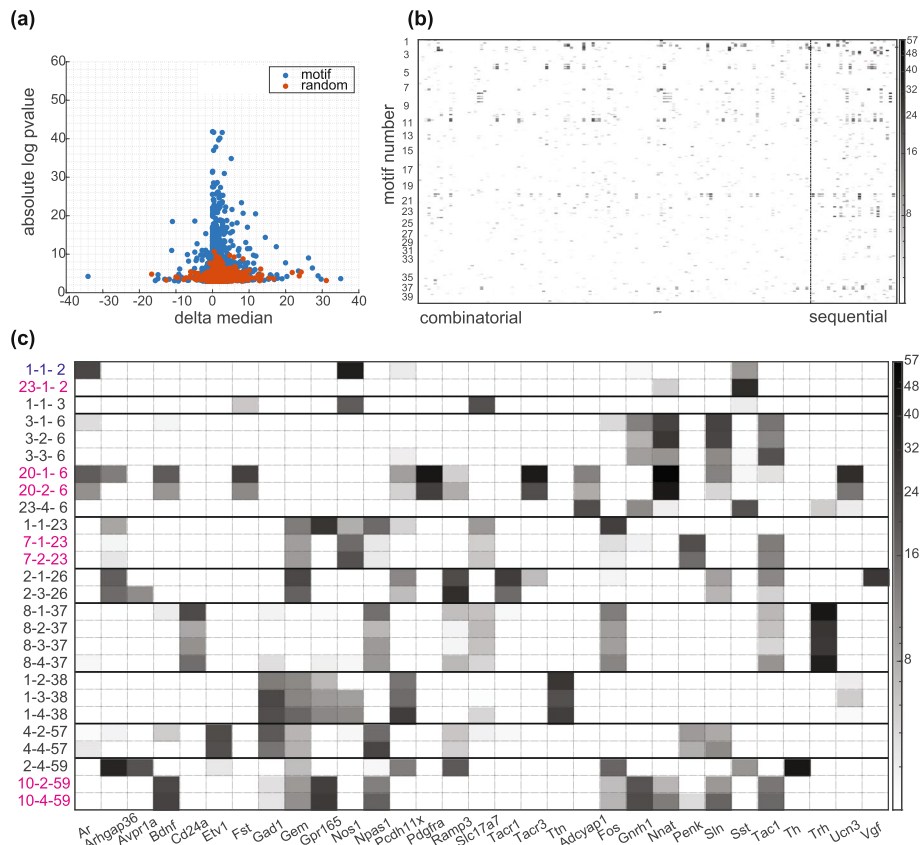
Fig. 7 Motif specific gene expression analysis for hypothalamus motifs. **a** Absolute log *p* value against difference in median gene expression in motif and non-motif cells. Blue markers show experimental results; red markers are the results for one random selection of cells. The expression values are not normalized. **b** Heatmap of absolute log *p* values for genes (on the horizontal axis) and motif cases (on the vertical axis). Out of 155 genes, 135 were measured by combinatorial FISH and 20 were measured in sequential rounds of FISH. These two groups are separated with a dashed line. **c** Selected cases and genes with at least one absolute log *p* value greater than 20. The motif cases are sorted by cell type. The gene names and their motif address (motif number-position-cell type) are included on the *x*- and *y*-axis, respectively

FISH rounds, typically genes with higher expression levels, exhibit a higher prevalence of significant cases in this analysis. This difference is probably related to the fact that gene expression values for sequential genes are generally under dispersed (Additional file 1: Fig. S6), resulting in the upregulated or down regulated genes being more significant. Genes with absolute log *p* values greater than 20 are highlighted in Fig. 7c. In most cases, the significance of differential gene expression varies depending on the position in the motif. For example, Vgf is upregulated in motif 2, position 1, cell type 26 (i.e., 2-1-26 case). But its differential expression is not significant in position 3 of the same motif. This may indicate that within a spatial motif, cells of the same type in different positions can have differences in their gene expression and potentially distinct roles.

Identifying spatial motifs in a 3D sample

So far, spatial transcriptomics has mostly been applied to thin tissue sections or monolayer cultured cells. In these cases, obtained measurements are typically projected on a two-dimensional plane. However, recent advances have made it possible to map gene

expression in thicker tissue slices, resulting in 3D datasets [45–48]. This is an important step in development of spatial methods because it enables profiling of cells in their native context, which in many tissues of interest is inherently three-dimensional. Since our method operates on a neighborhood graph, it should be able to identify spatial motifs in 3D datasets as well. To test this, we applied SMORE on the cellular map of a 200-µm-thick slice of mouse anterior hypothalamus, encompassing over 78,000 cells [49]. The cells in this dataset are classified into 21 excitatory neuronal clusters, 26 inhibitory neuronal clusters, and 7 non-neuronal cell subclasses based on the expression of 156 genes (Fig. 8a).

Figure 8b illustrates the top 20 output motif nodes highlighted within the graph. Comparing this figure with the equivalent view from Fig. 8a (lower left) indicates the similarity of the highlighted motifs with the visual structure of tissue and agrees with our expectation that global shuffling is able to derive discernable patterns, along with other motifs that are less obvious. Figure 8c highlights the first 5 motifs along with the logos representing the identified motifs. Overall, our results demonstrate that SMORE can be used for analysis of both 2D and 3D spatial data.

The cells within the 3D graph of tissue structure tend to have a greater number of neighbors on average compared to those in 2D datasets. For instance, while the average
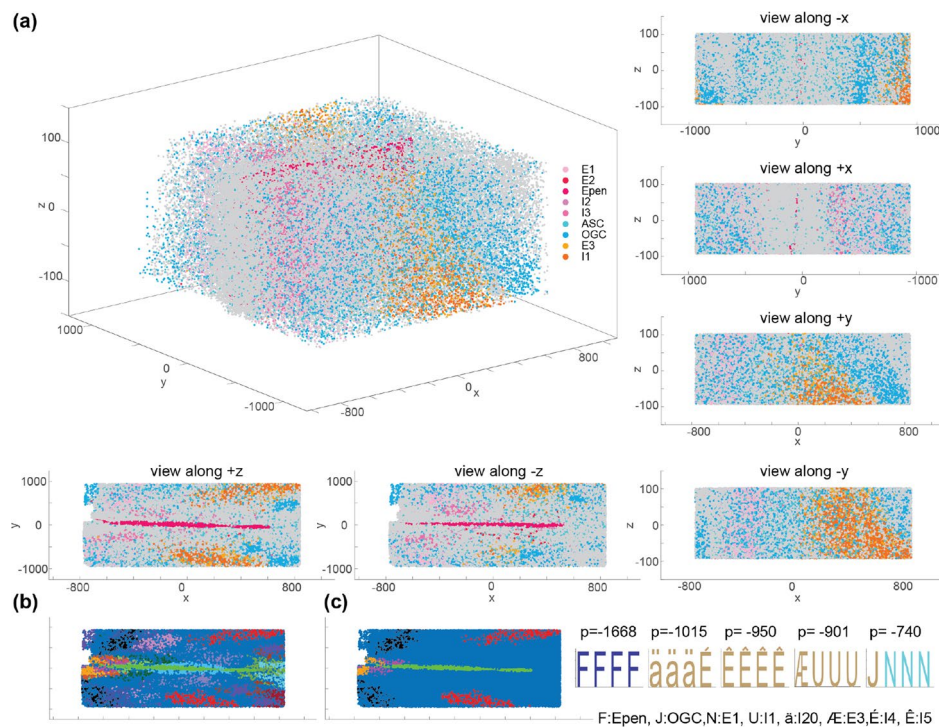

Fig. 8 Spatial motif analysis of a 200 µm slice of mouse hypothalamus. **a** The graph of 3D MERFISH dataset with classified cell subtypes shown as dots colored by subtype. The perpendicular views from different angles are shown alongside the 3D view for the primary graph. Epen, ependymal cells; ASC, astrocytes; OGC, oligodendrocytes. Excitatory and inhibitory subtypes started with E and I letter, respectively. **b** The first 20 spatial motifs obtained using global shuffling to generate the control data. **c** The first 5 output motifs along with their highlighted nodes on the tissue graph, and their respective log *p* values. Annotations are the same as Fig. 6 for the shared cell types

number of neighbors for the 2D mouse hypothalamus dataset in Fig. 6a is 5.9, it is 15.2 for the 3D hypothalamus dataset. Additionally, the 3D dataset comprises more than twice the number of cells in the 2D hypothalamus dataset (78,229 compared to 28,866). Consequently, the number of paths within the 3D dataset is substantially higher. Here, we used URPEN to sample 10% of the radial paths in this 3D tissue graph. Despite this downsampling, the number of radial paths for the 3D dataset was 1,838,093, while it was 696,141 for the specific tissue sections analyzed in Fig. 6, which were not downsampled. Due to the larger sample size, the log p values for the 3D dataset exhibit greater significance.

To further demonstrate scalability of our approach, we applied SMORE on a spatial atlas of the whole mouse brain (Allen Brain Cell Atlas) [50], which includes about 4 million cells profiled by MERFISH (Additional file 1: Fig. S7).

Spatial motifs can serve as structural signatures for tissue classification

Thus far, we have focused on spatial maps of cell types in neural tissues obtained through imaging-based methods. The complexity of neural tissues, with numerous intermingled cell types, and the single-cell resolution of imaging-based techniques make these datasets particularly well-suited for spatial motif analysis. However, SMORE is a general framework, agnostic of the methodology used for spatial profiling. Here, we demonstrate the broader utility of our method by analyzing transcriptomic maps of mouse embryos at embryonic day (E) 8.5 and 9, generated using Slide-seq [51].

Our analysis includes data from two E8.5 embryos, with 15 and 17 sagittal sections collected at 30 μm intervals, and one E9 embryo with 26 sagittal sections collected at 20 μm intervals. In total, the dataset encompasses 256,487 cells with gene expression profiles spanning 27,554 genes. Following the original study, we used 29 cell states as labels, each assigned by computationally mapping beads to a pre-existing single-cell reference. This approach effectively mapped cell states to their expected spatial domains, as shown for 10 selected cell states Fig. 9a.

As expected, due to the coarse-grained classification of cell states, most of the top motifs were composed of repeating patterns of a single cell state (Fig. 9b). These homotypic motifs represent regions of the sample that are broadly labeled, for example as brain or heart. Among the top 10 motifs, some also represent the boundary of two tissues, for example presomitic mesoderm and neuromesodermal progenitors (Fig. 9b, motif #10). Although these motifs lack the complexity seen in neural tissues, they still reveal statistically significant patterns. Therefore, we asked if motif frequency could be used to cluster structurally similar samples. Indeed, clustering based on the frequency of the top 30 motifs successfully grouped the two E8.5 samples together, separating them from the E9 embryo (Fig. 9c).

An advantage of sequencing-based spatial transcriptomics methods, like Slide-seq, is their ability to capture a comprehensive, unbiased view of gene expression across the tissue. We performed differential gene expression analysis, comparing cells within motifs with the cells of the same type elsewhere in the samples, and found numerous highly significant cases (Fig. 10a, b). A notable example is the lower expression of retinoic acid pathway members Crabp1 and Crabp2 in brain cells positioned within a sequence of anterior neuroectoderm cells (Motif #25, Fig. 10c, d). This observation aligns with
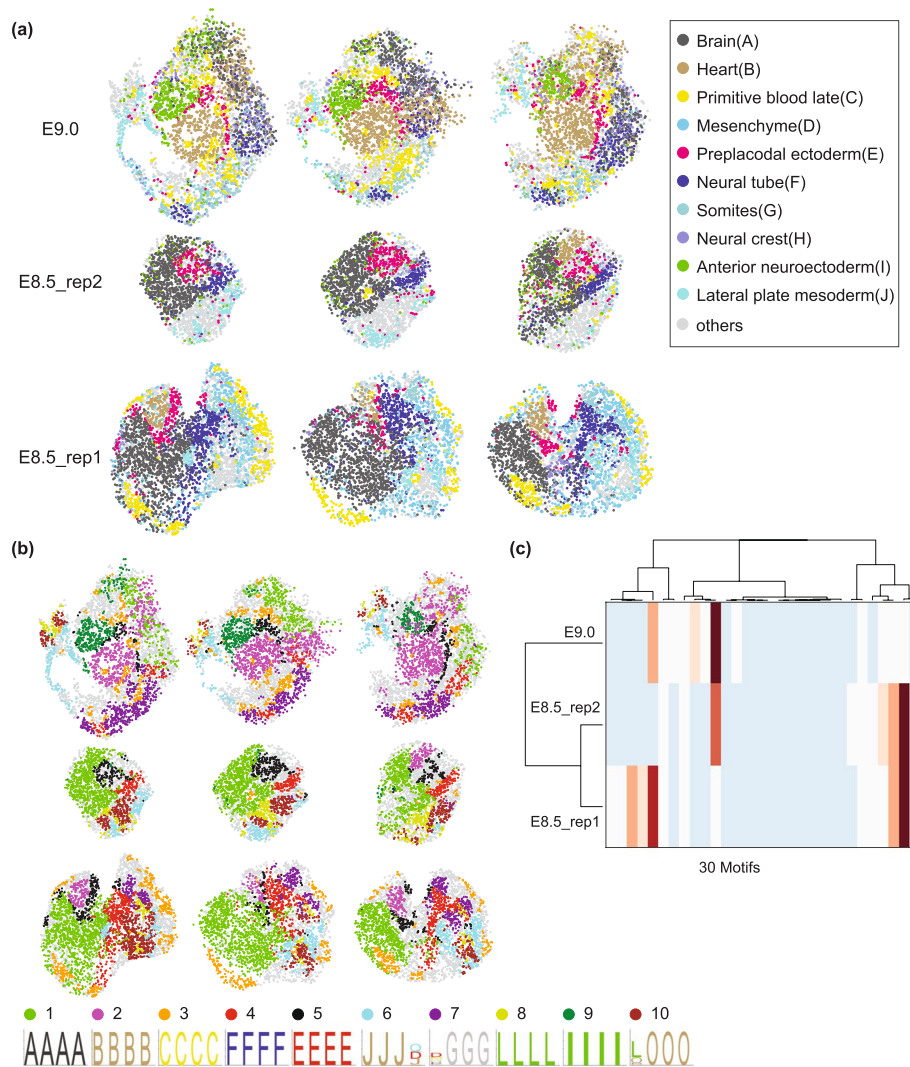
Fig. 9 Spatial motif analysis of a slide-seq mouse embryo dataset. **a** Cell type distributions in two E8.5 embryos (replicates) and one E9.0 mouse embryo. Three example sagittal sections from each embryo are shown. **b** First 10 output motifs obtained using global shuffling to generate the control data are highlighted on the tissue graph. Obtained motifs are represented by logos in the bottom. Annotations for the cell types involved in the motifs are either listed in the legend for panel (**a**) or at the bottom of panel **b**. Each motif is indicated by a different color in the highlighted tissue graph, with the colors noted on top of the motif logos. **c** Clustergram obtained based on correlation of the frequency of first 30 motifs in different embryo tissues

previously reported anterior-posterior expression domains of these genes [52] and may suggest region-specific modulation of RA signaling. Overall, among the 181 unique significant genes (log(p)< -80 and absolute delta median > 0.5), 40 overlapped with the 352 top enriched genes along the anteroposterior and dorsoventral axes reported in the original study [51]. This overlap, observed from a pool of 27,554 genes, is statistically significant (p-value $= 8.99e - 38$).

Together, this analysis demonstrates SMORE's versatility in identifying spatial motifs across diverse tissue types and transcriptomics platforms. It also highlights the scalability of our motif-specific gene expression analysis for genome-wide, sequencing-based data and showcases clustering of tissues based on frequency of their spatial motifs.
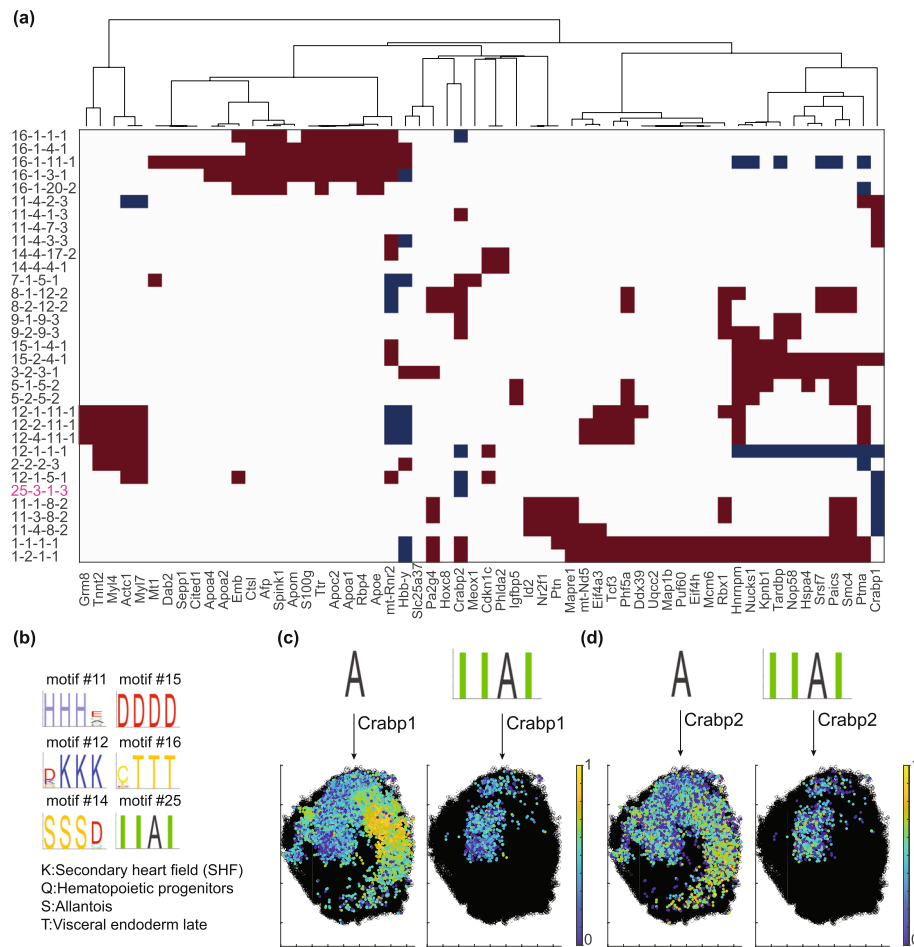
Fig. 10 Motif specific gene expression analysis of mouse embryonic samples. **a** Selected cases and genes with at least one absolute log *p*-value greater than 100 and absolute delta median greater than 1. The gene names and their motif address (motif number - position in motif - cell type - embryo number; MPCE) are included on the *x*- and *y*-axis, respectively. Embryos are numbered as 1:E8.5_rep1,2 : E8.5_rep2, and3 : E9. **b** Motif logos for the motifs that are not included in Fig. 9 but are present in the panel (**a**) heatmap. Cell type annotations are shown at the bottom. **c** Expression level of Crabp1 in all "brain" cells (cell state A) in the E9.0 embryo viewed from the top along the *z*-axis (left) compared to brain cells in motif #25 (right). **d** same as **c** for Crabp2 expression level. Both Crabp1 and Crabp2 are identified as significantly downregulated in brain cells within motif #25 compared to other brain cells, as highlighted in the panel (**a**) heatmap

Discussion

Spatial transcriptomics is a rapidly growing field. We have seen significant innovation in the field over the past few years, resulting in a multitude of techniques and consistent improvements in their efficiency and scalability. Concurrently, application of spatial transcriptomics has expanded beyond specialized groups to the broader community of biomedical researchers. There are already several commercial platforms available to researchers, and it is expected that additional options will become available in the near future. Therefore, the need for innovative computational methods to extract biologically relevant information from this type of data is on the rise.

In this work, we introduce a method for identifying patterns of cell type arrangements with arbitrary length. There are two major contributions that make this task

possible: a method for unbiased sampling of paths from a graph (URPEN) and a method for identifying motifs in such samples (SMORE). Our approach is general and can be applied to any system with sufficient complexity profiled with any spatial omics method. This includes solid tumors and organoids, where greater heterogeneity increases the need for statistical analysis. It is also not limited to two-dimensional maps and can be readily adapted for three-dimensional data. In addition to the application presented here, individual components of our method can be independently employed in a range of other contexts that are modeled as a graph. For example, path sampling via random walk has been used in applications ranging from network representation learning [53] to estimation of similarity measures [54]. Unbiased path sampling using URPEN can offer benefits over random walk in such applications. Detailed evaluation of performance improvement with URPEN needs further investigation.

Spatial motifs can be explained in terms of their functional significance or developmental mechanisms that generate a specific cell type arrangement. Therefore, they can be used to generate hypotheses and further our understanding of tissue biology. We have provided a few examples in this study. This includes association of RBCs with type 2 and 6 cone bipolar cells that can be involved in the primary scotopic pathway and downregulation of Grm6 in type 5b ON bipolar cells near type 1b cells that offers a possible mechanism for the function of these atypical bipolar cells. Interpretation of each spatial motif at this point can only be done on a case-by-case basis, in the context of what is known about the cell types involved. This can begin with length 2 motifs, which are easier to interpret, and progress to higher lengths in a stepwise manner. For motifs that consist of more than one seed, each seed can also be investigated separately.

Our findings also underscore the potential of spatial motifs as powerful tools for systematically characterizing tissues based on their cellular architecture. It will be interesting to explore more systematic ways of utilizing spatial motifs for characterizing tissues. For example, the set of all spatial motifs in a sample can provide a quantitative representation of the tissue structure. These representations can be valuable in classifying tissues with subtle differences in their cellular architecture, such as various cancer subtypes. In this study, we showcase an example of such classification for mouse embryonic tissues from E8.5 and E9 stages. To enhance this utility, further analysis is required to optimize key factors, including the normalization of high- and low-frequency motifs, selection between motifs, incorporation of features beyond motif frequency, and identification of the optimal motif length for best performance.

Our analysis here is constrained by certain technical limitations of the existing data. In datasets we analyzed, each cell is represented by a point in the tissue, which is typically the center of its nucleus. This ignores variation in size of the cells and their receptive fields. Position of the cell body may also not be a good indicator of neuronal connectivity. Furthermore, our motifs are currently confined to short range local neighborhoods surrounding each cell. Expanding the method to incorporate longer range interactions could be an interesting next step, either by clustering identified motifs or by permitting gaps in the motif sequence. Gene expression analysis can be informative as shown here. However, gene panels in imaging based spatial transcriptomics datasets are often selective, focusing on known cell type markers. In contrast, sequencing-based spatial methods offer expansive genomic coverage, but their measurements can be more noisy and

sparse. As more, scalable and multimodal spatial methods become available, SMORE has the potential to discover more intricate relationships between the spatial positioning of cells and their functional characteristics, including gene expression.

Conclusions

In conclusion, SMORE provides a novel and robust method for identifying spatial motifs in complex tissue structures. By capturing the sequential order of cells, SMORE fills a critical gap and complements the rapidly growing suite of spatial analysis methods. We present substantial algorithmic developments that enable efficient, unbiased sampling of paths from neighborhood graphs and discovery of motifs in the resulting cell type sequences. By rigorously investigating the performance of each component of our method, we quantitatively demonstrate their accuracy, specificity, and sensitivity. The application of SMORE to spatial maps of the mouse retina, brain, and embryonic tissue illustrates its utility in uncovering previously unrecognized patterns of cell type organization and contextualizing their biological significance through gene expression analysis. We further demonstrate the versatility and scalability of our method by analyzing samples spanning a wide range of cell and cell type counts, profiled using different spatial transcriptomics platforms. Together, our results highlight the capability of SMORE to illuminate the substantial complexity of neural tissues, provide novel insight even in well studied models, and generate experimentally testable hypotheses.

Methods

In order to apply the SMORE method on the spatial structure of the cell types, spatial transcriptomics dataset is imported and the neighborhood graph based on cell positions is created. Control data is generated using one of two methods: global shuffling and kernel shuffling. As an example, Additional file 1: Fig. S1 illustrates a kernel for the center node (highlighted in magenta). In this figure, K is set to 4 and 6, indicating that all nodes within K neighborhoods of the center node are part of the kernel.

nTrain instances of shuffled labels are generated. Labels for fixed nodes are not shuffled and labels for the other nodes are not shuffled with the fixed nodes. The graph is sampled with URPEN with the specified sampling frequency and the labels for the sampled paths are imported from either the original cell type labels or the set of nTrain shuffled labels created for the control data. After incorporating the reverse paths into the dataset, both primary and control data are fed into the SMORE method to identify motifs. The algorithms for path sampling and motif discovery are described in detail below.

Uniform path sampling

The topological relations inside the spatial data can be represented as graphs. A graph is defined as an ordered pair $G = (V;E)$ consisting of a nonempty set of vertices V and a set of edges E of two-element subsets of V. In the following, we are dealing with undirected graphs, but the path sampling algorithm can equally be applied to directed graphs as well. Vertices in V are assumed to be uniquely indexed by the integers $1, \ldots, n$, where $n = |V|$ is defined as the size of the graph. $v > u$ is used to indicate that the index of a vertex v is larger than that of a vertex u. For a vertex $v \in V_0$, where V_0 is a subset of vertices, its forward neighborhood with respect to

$V_0, N_{frw}(v, V_0)$, is defined as the set of all vertices from $V \setminus V_0$ which are adjacent to v. The neighborhood of a vertex is simply its forward neighborhood with respect to the empty set, $\emptyset$.

The developed algorithm for finding motifs in graphs consists of two main components; first a method to uniformly sample the graph, and second, a procedure to find motifs inside obtained sequences. The sampling algorithm takes a graph G and all paths inside the graph are sampled uniformly. In a network, a path is a walk that does not intersect itself. Selected paths are also constrained to be radial. Radial condition in a spatially embedded network is defined as the requirement that actual physical distance along a path monotonically increases along the sequence of edges in the path. The graph on spatial dataset is created based on the distance between nodes, therefore, the radial requirement enables us to better interpret output motifs in our experiment on the real dataset.

Algorithm 1 Path enumeration (G, k)

Input: A graph $G = (V, E)$ and an integer $1 \le k \le |V|$
Output: All length-k paths in G
1: **for** each vertex $\nu \in V$ **do**
2:     $V_{Forward} \Leftarrow (u \in N_{frw}(\{\nu\}, \emptyset)$
3:     **Call** ExtendPath$(\{\nu\}, V_{Forward})$
4: **end for**

ExtendPath$(V_{Path}, V_{Forward})$

E_1:   **If** $|V_{Path}| == k$ and $V_{Path}(k) > V_{Path}(1)$
E_2:   **Return** V_{Path}
E_3:   **While** $V_{Forward} \ne \emptyset$
E_4:   Remove an arbitrary chosen vertex w from $V_{Forward}$
E_5: $V'_{Forward} \Leftarrow \{u \in N_{frw}(w, V_{Path})$
E_6:   **Call** ExtendPath$(V_{Path}, V'_{Forward})$

The Path Enumeration algorithm (PEN) (Algorithm 1) enumerates all paths of length k within the graph. The algorithm begins with a vertex v from the input graph and adds only those vertices to the set that are neighboring the newly added vertex w but are not already in V_{path}. To prevent enumeration of both the path and its reverse, the index of the last vertex in the enumerated path must be greater than that of v, though this requirement is not necessary for directed graphs. The proof for the correctness of PEN is similar to that of the ESU algorithm [22].

We can modify the PEN algorithm to enumerate a subset of paths such that each path is reached with equal probability. This is implemented by calling the Extend-Path function at lines 3, and E_6 of the PEN algorithm with probability p_d. This new algorithm is called Uniform Random Path Enumeration, URPEN. The method is tested in the "URPEN enables efficient and unbiased sampling of neighborhood graphs" section on a random graph to validate its accuracy numerically.

Samadi *et al. Genome Biology*   (2025) 26:3

Page 25 of 30

Compute significance

The significance of each initial seed is obtained by the negative binomial test. The Poisson distribution arises naturally in the study of data taking the form of counts. If a data point y follows the Poisson distribution with rate θ, then the probability distribution of a single observation y is $y \sim \mathrm{Poisson}(\theta)$. The Poisson model for data points $\mathbf{y_v} = [y_1, y_2, \ldots, y_n]$ can be extended to the form $y_i \sim \mathrm{Poisson}(w_i \theta)$, where the w_i values are known positive explanatory values proportional to the population, and θ is the unknown parameter of interest. Seed number in each graph can be modeled as a Poisson distribution where y_i is the number of paths of some specific type (seed) in the graph, and θ is the underlying rate in units of seeds per graph.

To perform Bayesian inference, we need a prior distribution for the unknown rate. We use a gamma distribution as prior, which is conjugate to the Poisson. With prior distribution $\mathrm{Gamma}(\alpha, \beta)$, the resulting posterior distribution is obtained as

$$\theta|\mathbf{y_v} \sim \mathrm{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + \sum_{i=1}^{n} w_i\right) \tag{1}$$

The known form of the prior and posterior densities can be used to find the marginal distribution for a single observation, which has a predictive distribution as

$$p(y_0|\mathbf{y_v}) = \frac{p(y_0|\theta)p(\theta|\mathbf{y_v})}{p(\theta|y_0, \mathbf{y_v})} = \mathrm{NegBin}(\alpha_n, p_n), \tag{2}$$

where $\alpha_n = \alpha + \sum_{i=1}^{n} y_i$, and $p_n = \frac{w_0}{(\beta_n + w_0)}$ with $\beta_n = \beta + \sum_{i=1}^{n} w_i$. Assuming that y_0 is the seed number in the primary graph, p value for some specific observation is obtained as

$$P = \sum_{k=y_0}^{\infty} \binom{n+k-1}{k} p_n^k (1-p_n)^n, \tag{3}$$

where y_0, and $y_i, i = 1, 2, \ldots, n$ are the number of ZNIC sites of the specific seed in primary and control data, respectively. w_i numbers are the total number of ZNIC seeds in the respective graph. In our experiments, prior α and β values are assumed to be equal to y_0, and w_0, respectively. The first nEval significant seeds according to this criterion are passed to the next stage of refinement and enrichment.

Refinement and nested seed enrichment

The refinement and seed enrichment both use the same process of enrichment, except that refinement is only one iteration, and seed enrichment is NREFIter iterations. nEval motifs from initial evaluation step are first enriched for one iteration and top NREF motifs are further refined in seed enrichment block for NREFIter iterations or until p value does not improve. At each iteration, all seeds with positive likelihood ratio scores with respect to the PWM matrix obtained from the previous iteration are sorted with their score and p values obtained in the initial evaluation block and their ZNIC counts are computed. More specifically, it's computed how many ZNIC samples each seed contributes to the previous samples. These counts are used to obtain significance with

the same negative binomial test obtained in 4. For the negative control data, sum of the counts over nTrain control data are used for significance computation.

The PWM score that minimizes p value (maximizes absolute log p value) is selected to create the PWM matrix for the next iteration. There is an option in the algorithm to use differential enrichment where seeds are added to the PWM until the p value is decreasing. For example, if there are four PWM score thresholds with ZNIC log p values, $[-10, -11, -9, -13]$, the default mode will consider lowest score threshold which is equivalent to log $p$value $= -13$, but differential p value will consider the threshold corresponding to log $-p$value $= -11$. The differential p value option will generally lead to a simpler motif structure. Default mode is used for bipolar and 3D hypothalamus dataset, and differential p value is used for the preoptic area of mouse hypothalamus.

Maximum likelihood estimation is used to estimate a new version of the motif PWM matrix for the next iteration. This step is iterated until the p value is decreasing or the maximum number of NREFIter is performed. In order to perform maximum likelihood estimation, let us assume that we have L distinct cell types in our dataset, encoded as integer numbers from 1 to L. Assuming that the initial seed consists of W letters, $S = s_1, s_2, \ldots, s_W$, the PWM matrix, M, is a $L \times W$ matrix where W is the length of the searched motif. Elements of the PWM matrix in the first iteration are obtained as follows,

$$\mathbf{M}(i,j) = \begin{cases} 1 + \rho \mathbf{b}(s_j)/1 + \rho & \text{if } s_j = i \\ \rho/1 + \rho & \text{otherwise} \end{cases} \tag{4}$$

where ρ is the Dirichlet prior set to 0.01 in our experiment, and b is the $L \times 1$ vector of background probabilities of the cell types. For the subsequent iterations, the maximum likelihood estimation for M matrix is obtained as follows [23],

$$\hat{\mathbf{M}} = \sum_{i=1}^{K} P_i Z_i / N_i \mathbf{I}_i + \rho \mathbf{b} \tag{5}$$

The PWM matrix, $\mathbf{M}$, is obtained by normalizing $\hat{\mathbf{M}}$ through the columns. K is the number of seeds involved in the motif, and $\mathbf{I}_i$ is the indicator matrix of ith seed, which is equivalent to PWM matrix of 4 with ρ set to 0. Zi, Ni, and Pi are the incremental ZNIC counts, total ZNIC counts, and log p value of the ith seed involved in the motif, respectively. Total ZNIC counts are the number ZNIC sites of the ith seed and incremental ZNIC counts are the fraction of these sites that don't have any node in common with previous seeds, up to i'th seed.

Applying SMORE on synthetic and real data

We have tested SMORE on synthetic data and real data. The results on synthetic data helps to quantify the algorithm's performance by assessing its ability to identify known truth patterns at various embedding frequencies. Subsequently, SMORE has been applied to multiple real spatial transcriptomic datasets, each of which is thoroughly detailed in its corresponding section.

The background frequency in synthetic data experiment is assumed to be near uniform, with

$$\mathbf{b}_F = [1, 1/2, 1/3, \ldots, 1/12]^{1/4}, \tag{6}$$

where $\mathbf{b}_F$ is the background frequency and it is assumed to be known to the algorithm. Cell type labels, other than the ones for the motifs, are distributed randomly among available nodes according to their frequency. This specific form of background frequency is arbitrary and is designed to have at most twofold difference in the cell type frequency to avoid potential unwanted repeating cell type patterns like AAAA.

Output motifs are represented by sequence logos. Some output motifs are simple in structure and some are more complex, consisting of multiple cell types in each position. Simple logos like AAAA indicate a repetitive pattern of A type cells interconnected in the graph within that particular tissue. In more complex motifs like (AB)AAA, the first position can be occupied by either A or B. Embedded length-4 patterns in the synthetic data follow a format like (A/B)CDE, where one variable position can be filled by either cell type A or B, each with equal probability. The remaining positions in the pattern (CDE) are fixed to specific cell types. For a 2% embedding, the nodes corresponding to 60 random path samples from the total length-4 sampled paths are labeled as ACDE and another 60 as BCDE, while all other nodes in the graph are randomly labeled.

In Fig. 3a, the Pearson correlation coefficient (PCC) between the extracted motifs and the embedded motifs is shown. The Tomtom method [55] is employed to identify the best enriched matches among the 10 output motifs from SMORE. Tomtom searches for the best match by considering all possible shifts of the query motif with respect to the target motif (the embedded motif in this case). The matching position weight matrices (PWMs) are aligned using the obtained offsets and overlaps from Tomtom, and the PCC is computed and plotted to evaluate the pipeline's accuracy. Each embedded motif word corresponds to a 12×4 PWM matrix, employing a Dirichlet prior with a weight of 0.01. The Dirichlet prior is a uniform distribution of the cell types.

Computing TPR and FPR

TPR is computed as $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, where TP (true positive) denotes the number of output motifs with a correlation exceeding the threshold (0.95 in our context) and a p value below the specified threshold. FN (false negative) is the number of instances with a correlation surpassing the threshold but a p value greater than the specified threshold. Thus, $\mathrm{TP} + \mathrm{FN}$ represents the overall number of cases with a correlation exceeding the threshold.

FPR is defined as $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$, with FP (false positive) indicating the number of output motifs possessing a correlation below the threshold and a p value below the specified threshold. TN (true negative) is the number of output motifs with a correlation below the threshold and a p value exceeding the specified threshold. Consequently, $\mathrm{FP} + \mathrm{TN}$ denotes the total number of cases with a correlation below the threshold.

Gene expression analysis

One potential method for deriving a functional interpretation of these motifs involves examining gene expression disparities between the cells participating in the motif and those that are not involved. This analysis is carried out in the "Cells within spatial motifs exhibit gene expression differences compared to other cells of the same type"

section. Each motif is composed of multiple positions, and each position includes one or more cell types. For example, the motif (AB)CDE comprises cell types A and B in its first position. Among all cells labeled as A (N_A), only a subset, N_{AM}, takes part in the initial position of this particular motif. Each cell has a gene expression profile. For each gene, the delta median(dMedian) is calculated by subtracting the median expression of the gene in the subset N_{AM} cells involved in the motif from the median expression of that gene across all N_A cells of the specific type. The significance for this delta median value is then computed against a random selection of N_{AM} cells from the total N_A cells of that type.

Assume that there are N_A cell types of A, N_{AM} of these cell types is in motif (AB) CDE, with dMedian expression x_0. One sided p value (significance) for these cells is $p(\text{dMedian} \geq x_0)$. The probability that dMedian expression of these N_{AM} cells is larger than x_0 is the probability that at least $N_0 = \text{floor}(N_{AM}/2 + 1)$ of these cells have expression greater than x_0,

$$p_0 = \frac{\sum_{k=N_0}^{N_{AM}} \binom{N_H}{k}\binom{N_L}{N_{AM}-k}}{\binom{N_A}{N_{AM}}} \tag{7}$$

Where N_L is the number of cells (out of all N_A cells) that have lower than x_0 delta expression, and N_H is the number of cells that have higher than x_0 delta expression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03467-5.

Additional file 1: Supplementary Material for SMORE: spatial motifs reveal patterns in cellular architecture of complex tissues. This file contains recommendations for handling large sample numbers, side-by-side comparison of SMORE with HistoCAT and ImaCytE's performance on the retinal bipolar dataset, and analysis of a whole mouse brain spatial transcriptomics atlas, as well as Figs. S1 to S7 and Table S1.

Additional file 2: Review history.

Declarations

References

1. Moses L, Pachter L. Museum of spatial transcriptomics. Nat Methods. 2022;19(5):534–46.
2. Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. Nat Biotechnol. 2023;41(6):773–82.
3. Li Y, Lac L, Liu Q, Hu P. ST-CellSeg: cell segmentation for imaging-based spatial transcriptomics using multi-scale manifold learning. PLOS Comput Biol. 2024;20(6):e1012254.
4. Elhanani O, Ben-Uri R, Keren L. Spatial profiling technologies illuminate the tumor microenvironment. Cancer Cell. 2023;41(3):404–20.
5. Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022;19(2):171–8.
6. Getis A, Ord JK. The analysis of spatial association by use of distance statistics. Geogr Anal. 1992;24(3):189–206.
7. Ripley BD. The second-order analysis of stationary point processes. J Appl Probab. 1976;13(2):255–66.
8. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VR, Schulz D, et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. Nat Methods. 2017;14(9):873–6.
9. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018;174(4):968–81.
10. Väyrynen SA, Zhang J, Yuan C, Väyrynen JP, Dias Costa A, Williams H, et al. Composition, spatial characteristics, and prognostic significance of myeloid cell infiltration in pancreatic cancer. Clin Cancer Res. 2021;27(4):1069–81.
11. Aoki T, Chong LC, Takata K, Milne K, Hav M, Colombo A, et al. Single-cell transcriptome analysis reveals disease-defining T-cell subsets in the tumor microenvironment of classic Hodgkin lymphoma. Cancer Disc. 2020;10(3):406–21.
12. Nieto P, Elosua-Bayes M, Trincado JL, Marchese D, Massoni-Badosa R, Salvany M, et al. A single-cell tumor immune atlas for precision oncology. Genome Res. 2021;31(10):1913–26.
13. McCaffrey EF, Donato M, Keren L, Chen Z, Delmastro A, Fitzpatrick MB, et al. The immunoregulatory landscape of human tuberculosis granulomas. Nat Immunol. 2022;23(2):318–29.
14. Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell. 2020;182(5):1341–59.
15. Bhate SS, Barlow GL, Schürch CM, Nolan GP. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. Cell Syst. 2022;13(2):109–30.
16. Chen Z, Soifer I, Hilton H, Keren L, Jojic V. Modeling multiplexed images with spatial-LDA reveals novel tissue microenvironments. J Comput Biol. 2020;27(8):1204–18.
17. Jackson HW, Fischer JR, Zanotelli VR, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature. 2020;578(7796):615–20.
18. Danenberg E, Bardwell H, Zanotelli VR, Provenzano E, Chin SF, Rueda OM, et al. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. Nat Genet. 2022;54(5):660–9.
19. Somarakis A, Van Unen V, Koning F, Lelieveldt B, Höllt T. ImaCytE: visual exploration of cellular micro-environments for imaging mass cytometry data. IEEE Trans Vis Comput Graph. 2019;27(1):98–110.
20. Stoltzfus CR, Filipek J, Gern BH, Olin BE, Leal JM, Wu Y, et al. CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. Cell Rep. 2020;31(3):107523.
21. Polonsky M, Gerhardt LM, Yun J, Koppitch K, Colón KL, Amrhein H, et al. Spatial transcriptomics defines injury specific microenvironments and cellular interactions in kidney regeneration and disease. Nat Commun. 2024;15(1):7010.
22. Wernicke S. Efficient detection of network motifs. IEEE/ACM Trans Comput Biol Bioinforma. 2006;3(4):347–59.
23. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–40.
24. Schreiber F, Schwöbbermeyer H. Frequency concepts and pattern detection for the analysis of motifs in networks. In: Transactions on computational systems biology III. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. pp. 89–104.
25. Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics. 2004;20(11):1746–58.
26. West ER, Lapan SW, Lee C, Kajderowicz KM, Li X, Cepko CL. Spatiotemporal patterns of neuronal subtype genesis suggest hierarchical development of retinal diversity. Cell Rep. 2022;38(1):110191.
27. West ER, Cepko CL. Development and diversification of bipolar interneurons in the mammalian retina. Dev Biol. 2022;481:30–42.
28. Reese B, Necessary B, Tam P, Faulkner-Jones B, Tan SS. Clonal expansion and cell dispersion in the developing mouse retina. Eur J NeuroSci. 1999;11(8):2965–78.
29. Sharpe LT, Stockman A. Rod pathways: the importance of seeing nothing. Trends Neurosci. 1999;22(11):497–504.
30. Tsukamoto Y, Omi N. Functional allocation of synaptic contacts in microcircuits from rods via rod bipolar to AII amacrine cells in the mouse retina. J Comp Neurol. 2013;521(15):3541–55.

31. Tsukamoto Y, Omi N. Classification of mouse retinal bipolar cells: type-specific connectivity with special reference to rod-driven AII amacrine pathways. Front Neuroanat. 2017;11:92.
32. Edsgärd D, Johnsson P, Sandberg R. Identification of spatial expression trends in single-cell gene expression data. Nat Methods. 2018;15(5):339–42.
33. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. Nat Methods. 2018;15(5):343–6.
34. Sun S, Zhu J, Zhou X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nat Methods. 2020;17(2):193–200.
35. Hao M, Hua K, Zhang X. SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. Bioinformatics. 2021;37(23):4392–8.
36. Andersson A, Lundeberg J. sepal: identifying transcript profiles with spatial patterns by diffusion-based modeling. Bioinformatics. 2021;37(17):2644–50.
37. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, et al. SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods. 2021;18(11):1342–51.
38. Zhu J, Sun S, Zhou X. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biol. 2021;22(1):1–25.
39. Vardi N, Morigiwa K. ON cone bipolar cells in rat express the metabotropic receptor mGluR6. Vis Neurosci. 1997;14(4):789–94.
40. Dhingra A, Vardi N. mGlu receptors in the retina. Wiley Interdiscip Rev Membr Transp Signal. 2012;1(5):641–53.
41. Zeitz C, van Genderen M, Neidhardt J, Luhmann UF, Hoeben F, Forster U, et al. Mutations in GRM6 cause autosomal recessive congenital stationary night blindness with a distinctive scotopic 15-Hz flicker electroretinogram. Investig Ophthalmol Vis Sci. 2005;46(11):4328–35.
42. Della Santina L, Kuo SP, Yoshimatsu T, Okawa H, Suzuki SC, Hoon M, et al. Glutamatergic monopolar interneurons provide a novel pathway of excitation in the mouse retina. Curr Biol. 2016;26(15):2070–7.
43. Shekhar K, Lapan SW, Whitney IE, Tran NM, Macosko EZ, Kowalczyk M, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell. 2016;166(5):1308–23.
44. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science. 2018;362(6416):eaau5324.
45. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science. 2018;361(6400):eaat5691.
46. Alon S, Goodwin DR, Sinha A, Wassie AT, Chen F, Daugharthy ER, et al. Expansion sequencing: spatially precise in situ transcriptomics in intact biological systems. Science. 2021;371(6528):eaax2656.
47. Wang Y, Eddison M, Fleishman G, Weigert M, Xu S, Wang T, et al. EASI-FISH for thick tissue defines lateral hypothalamus spatio-molecular organization. Cell. 2021;184(26):6361–77.
48. Kuett L, Catena R, Özcan A, Plüss A, Schraml P, Moch H, et al. Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment. Nat Cancer. 2022;3(1):122–33.
49. Fang R, Halpern A, Rahman MM, Huang Z, Lei Z, Hell SJ, et al. Three-dimensional single-cell transcriptome imaging of thick tissues. Elife. 2024;12:RP90029.
50. Yao Z, van Velthoven CT, Kunst M, Zhang M, McMillen D, Lee C, et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature. 2023;624(7991):317–32.
51. Sampath Kumar A, Tian L, Bolondi A, Hernández AA, Stickels R, Kretzmer H, et al. Spatiotemporal transcriptomic maps of whole mouse embryos at the onset of organogenesis. Nat Genet. 2023;55(7):1176–85.
52. Lyn S, Giguère V. Localization of CRABP-I and CRABP-II mRNA in the early mouse embryo by whole-mount in situ hybridization: implications for teratogenesis and neural development. Dev Dyn. 1994;199(4):280–91.
53. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: Association for Computing Machinery; 2014. pp. 701–10.
54. Zhang J, Tang J, Ma C, Tong H, Jing Y, Li J. Panther: fast top-k similarity search on large networks. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery; 2015. pp. 1445–54.
55. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2):1–9.
56. Samadi Z. Spatial Motif Recognition. GitHub. https://github.com/zsamadi/SMORE. Accessed 20 Dec 2024.
57. Samadi Z. zsamadi/SMORE: v1.0.0. Zenodo. 2024. https://doi.org/10.5281/zenodo.14510210.
58. West E. Bipolar-Serial-SABER-FISH-Analysis [Dataset]. Zenodo. 2021. https://doi.org/10.5281/zenodo.5716532.
59. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Data from: Molecular, spatial and functional single-cell profiling of the hypothalamic preoptic region [Dataset]. Dryad. https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248. Accessed 11 Mar 2024.
60. Sampath Kumar A, Tian L, Meissner A, Chen F. Spatial transcriptomic maps of whole mouse embryos [Dataset]. CZI Single-Cell Biology. 2022. https://cellxgene.cziscience.com/collections/d74b6979-efba-47cd-990a-9d80ccf29055.
61. Yao Z, van Velthoven CT, Kunst M, Zhang M, McMillen D, Lee C, et al. MERFISH spatial transcriptomics dataset of a single adult mouse brain [Dataset]. Allen brain cell atlas. 2024. https://allen-brain-cell-atlas.s3.us-west-2.amazonaws.com/index.html#metadata/MERFISH-C57BL6J-638850/20241115/.

Publisher's Note