communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-026-09898-z

Smoothie: efficient inference and integration of spatial co-expression networks from denoised spatial transcriptomics data

Chase Holdener [1,2] & Iwijn De Vlaminck [1]

Finding correlations in spatial gene expression is fundamental in spatial transcriptomics, as co-expressed genes within a tissue are linked by regulation, function, pathway, or cell type. Yet, sparsity and noise in spatial transcriptomics data pose significant analytical challenges. Here, we introduce *Smoothie*, a pipeline that denoises spatial transcriptomics data with Gaussian smoothing and constructs and integrates genome-wide co-expression networks. Utilizing implicit and explicit parallelization, Smoothie scales to datasets exceeding 100 million spatially resolved spots with fast run times and low memory usage. We demonstrate how co-expression networks measured by Smoothie enable precise gene module detection, functional annotation of uncharacterized genes, linkage of gene expression to genome architecture, and multi-sample comparisons to assess stable or dynamic gene expression patterns across tissues, conditions, and time points. Overall, Smoothie provides a scalable and versatile framework for extracting deep biological insights from high-resolution spatial transcriptomics data.

Recent advances in high-resolution spatial transcriptomics enable profiling of millions of spatially resolved transcriptomes within a single tissue sample. Yet, analyzing these data remains challenging due to noise, data sparsity, and the fact that each "spot" can capture partial gene expression from multiple cells. A powerful strategy for interpreting complex spatial RNA data is to measure weighted spatial gene co-expression networks, which capture how gene pairs co-express across physical locations in the tissue. Weighted co-expression network analysis was first introduced on microarray bulk RNA data[1,2] and has been adapted for single cell RNA sequencing and low-resolution spatial RNA sequencing[3,4]. In spatial transcriptomics, analyzing spatial gene modules and their networks can identify novel gene associations, even among previously uncharacterized genes, discovering new gene functions and new gene annotations[5]. Further, the gene correlation network provides a straightforward way to summarize the co-expression patterns of any gene set of interest, such as a gene family, a gene pathway, cell communication genes, transcription factor genes, or more. As we show here, gene correlation networks provide an avenue for harmonizing and cross-analyzing data from different tissues, a task for which few substantive solutions exist. Critically, constructing gene correlation networks does not require prior knowledge of tissue biology or architecture, and does not require companion data (e.g., imaging or single-cell transcriptomic data) or specialized preprocessing such as cell segmentation, which can be difficult to perform accurately[6].

Given the importance of spatial gene correlation networks, it is essential to measure them accurately and efficiently. Here, we introduce and validate *Smoothie*, a pipeline for precise inference and integration of spatial gene correlation networks in high resolution spatial transcriptomics. Smoothie performs Gaussian smoothing on the data to address noise and sparsity, constructs genome-wide gene correlation networks with a hybrid soft-hard network thresholding strategy, and identifies gene modules with random walk graph-based clustering for downstream analyses. For multi-dataset scenarios, Smoothie supports cross-sample gene co-expression network inference as well as a second-order correlation analysis to rank gene patterns as stable or dynamic between samples. Smoothie can process datasets exceeding 100 million spatial spots and over 20,000 genes in about 1 hour on a standard high-performance CPU. Thus, Smoothie can handle large datasets from the newest spatial platforms.

We demonstrate how Smoothie identifies gene modules in both the adult mouse cerebellum and the developing mouse embryo, where modules correspond to tissue regions, known cell types, and local genome

[1]Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, USA. [2]Department of Computational Biology, Cornell University, Ithaca, NY, USA.
e-mail: vlaminck@cornell.edu

positioning. We further use Smoothie to associate over 50 previously uncharacterized or predicted mouse genes with known cell types and spatial gene programs, and we spatially characterize the *Slc* transporter superfamily in the mouse embryo. Finally, we show how second-order gene correlation analysis can detect differences in expression patterns across timepoints for both the erythroid and epidermis cell populations in mouse embryonic development (E13.5-E16.5) as well as the mural antral granulosa cell type during induced mouse ovulation. In conclusion, Smoothie offers an efficient and versatile approach for biological discovery in spatial transcriptomics data by leveraging spatial gene correlation networks.

Results

Smoothie precisely measures spatial gene correlations and finds gene modules

Smoothie implements a multi-step pipeline to output spatial gene modules in an interpretable gene network format (Fig. 1A). The input is a spatial gene count matrix X with N spatial spots and G genes. In the first step, Smoothie applies Gaussian smoothing to remove fine-scale noise and to address data sparsity. Specifically, Smoothie moves a 2D Gaussian distribution with standard deviation σ microns across the (x, y) locations in the dataset and generates smoothed expression values for each gene by taking a Gaussian weighted average of the nearby gene values (Methods). Next, Smoothie creates a pairwise gene correlation matrix R of size $G \times G$ by computing the Pearson correlation coefficient (PCC) between all pairs of genes in parallel. To construct the gene correlation network, Smoothie uses hybrid soft-hard network thresholding with a hard threshold PCC^* and soft thresholding power β. The hard threshold PCC^* first removes weak gene correlations, only retaining correlations $R_{ij} | R_{ij} > PCC^*$. Next, soft thresholding involves shifting edge weights from the interval $(PCC^*, 1]$ to $(0, 1]$ followed by raising the shifted values to a power $\beta \geq 1$. Random walk network clustering then identifies gene modules[7]. The use of hybrid soft-hard network thresholding puts more emphasis on stronger, more confident edge weights, leading to improved clustering results. Choices for PCC^* and β can vary depending on the researcher's goals, with the lower bound for hard cutoff PCC^* determined by considering a spatially shuffled version of X (Methods).

Major analyses from the Smoothie pipeline include identifying precise spatial gene modules, investigating individual genes through spatial correlations, connecting gene correlations with genome position, examining specific gene sets or families, and more (Fig. 1B, Supplementary Fig. 2)[8]. For multi-dataset scenarios, Smoothie also enables spatial gene module integration across datasets and provides a way to rank genes as having stable or dynamic spatial expression patterns across diverse conditions using second-order correlation analysis (Fig. 1C).

Smoothie identifies spatial gene modules in the adult mouse cerebellum

We first applied Smoothie to the Slide-seq Puck_180819_12 cerebellum dataset[9], which contains 32,701 10 μm-resolution spots (≥5 UMIs each) and 6942 genes (≥50 detected spots). After applying Gaussian smoothing with 30 μm Gaussian standard deviation, we detected 17,208 gene pairs with PCC > 0.35 among 1206 genes (Supplementary Data 1, Methods), and weighted random network walk clustering with soft power $\beta = 4$ identified 10 gene modules (Supplementary Data 2). Tightening the cutoff to PCC > 0.6 revealed 270 edges among 118 genes and highlighted substructure in the correlation network (Fig. 2A). For example, Purkinje cell marker genes *Calb1*, *Pvalb*, *Car8*, and *Pcp4* (module 3) grouped in a closely connected

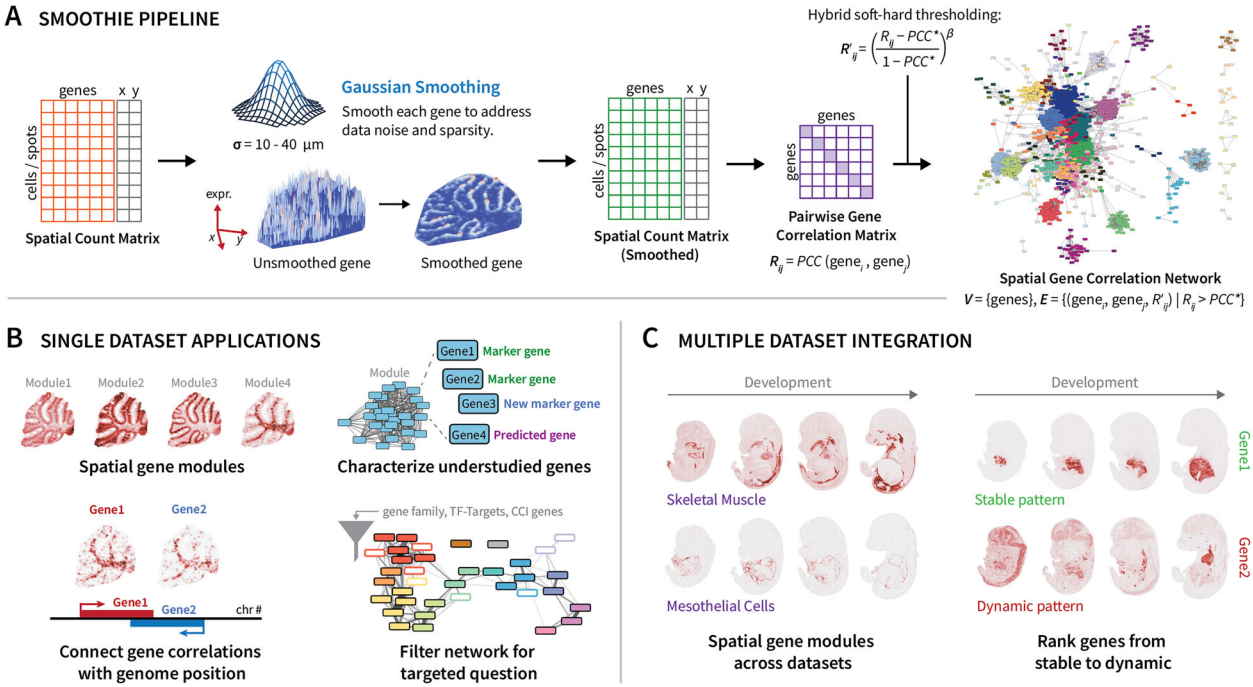

Fig. 1 | Overview of the Smoothie pipeline and its applications in one or many spatial datasets. A Smoothie takes a spatial count matrix as input and performs efficient Gaussian smoothing on each gene using a 2D Gaussian distribution with standard deviation σ microns. After calculating the correlation of every smoothed gene pair to create a gene correlation matrix R, the correlation matrix is converted into a gene network adjacency matrix using a hybrid soft-hard thresholding approach. A hard correlation threshold PCC^* removes weak gene correlations, and a soft thresholding power β puts higher emphasis on strong edge weights after linearly rescaling edge weights from the interval $(PCC^*, 1]$ to $(0, 1]$. The resulting network, with genes as vertices and weighted edges reflecting transformed spatial gene correlations, identifies spatial gene modules using random walk graph clustering (Methods). **B** Several powerful biological use cases follow from the gene correlation network. First, the network finds precise spatial gene modules grouped by pattern similarity. Second, gene modules characterize understudied or predicted genes by connecting them with known gene pathways, cell types, or tissue regions. Third, spatial gene correlations may reflect local genome positioning. Fourth, the gene correlation network may be filtered for any gene set of interest to investigate pattern diversity of the gene set members, such as a gene family, TFs and target genes, or cell-cell interaction (CCI) genes. **C** Smoothie provides integration analyses for multiple spatial dataset experiments, such as finding spatial gene modules across datasets as well as ranking genes as stable or dynamic across datasets using second-order gene correlations.

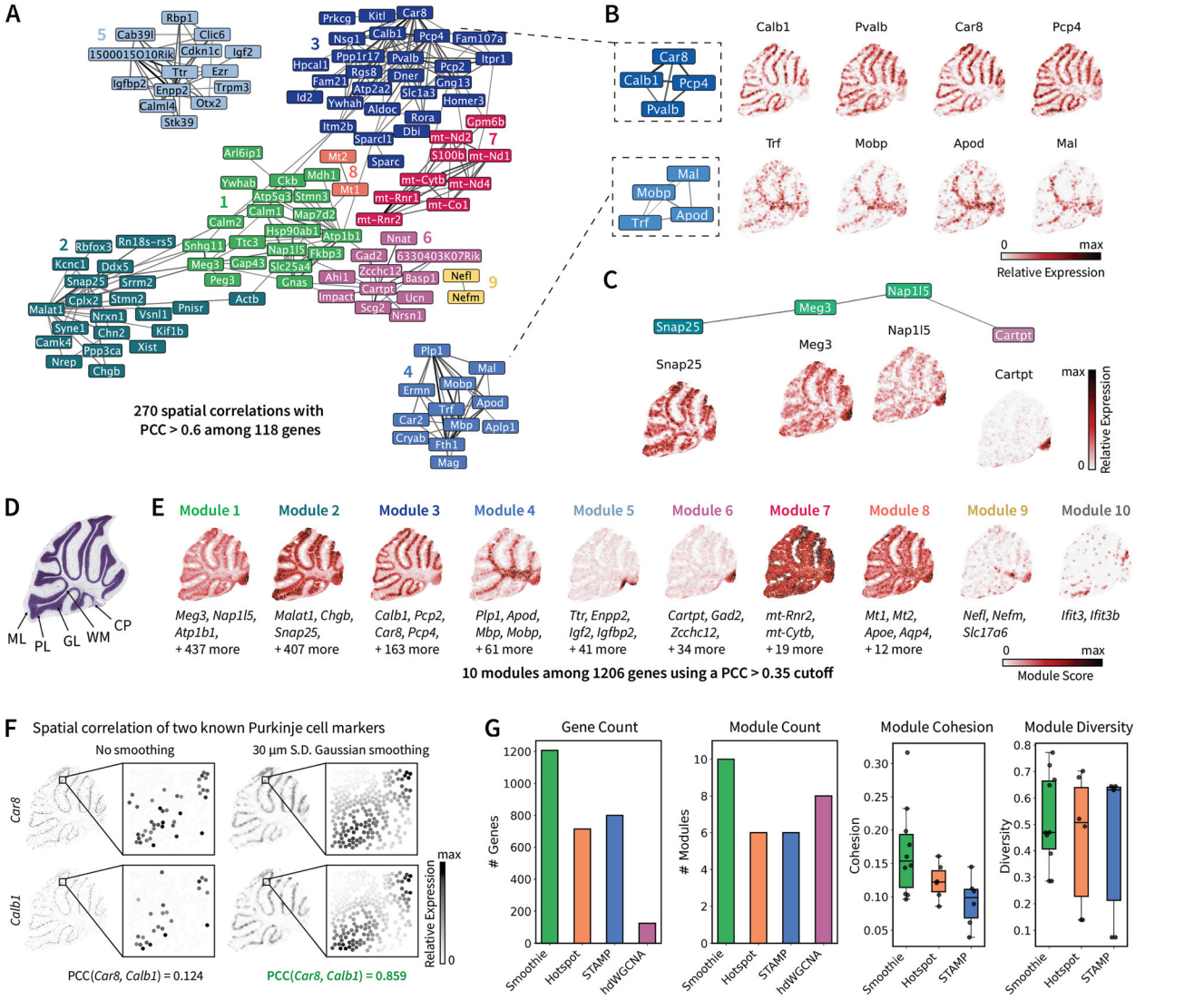

Fig. 2 | Smoothie identifies spatial gene modules in the adult mouse cerebellum. **A** The spatial gene correlation network, representing genes as nodes and spatial gene correlations as edges. Edges are weighted by the Pearson correlation coefficient (PCC) between the two gene surfaces, and the width and darkness of the line reflect edge weight. Nodes are colored by their assigned module. Network cutoff PCC > 0.6. **B** Example spatial expression patterns from tightly connected genes in the network. Plots show relative gene expression after normalization and smoothing. **C** Example spatial expression patterns from connected genes spanning multiple modules. **D** Nissl stained adult mouse cerebellum image from the Allen Institute's mouse.brain-map.org with labeled layers and regions. (ML = molecular layer, PL = Purkinje layer, GL = Granule layer, WM = white matter, CP = choroid plexus). **E** Module score plots for each of the 10 modules identified by Smoothie. The module score reflects average smoothed expression of all genes in the module after rescaling each gene's maximum value to 1 (Methods). **F** A zoomed in view of known Purkinje cell marker genes *Car8* and *Calb1* shows how Gaussian smoothing enhances the PCC between the two gene surfaces significantly. **G** Clustering benchmarks between Smoothie, Hotspot, STAMP, and hdWGCNA reveal that Smoothie identifies more modules with higher gene recall than other methods, while achieving higher module cohesion and similar module diversity levels to Hotspot and STAMP. hdWGCNA clusters much fewer genes compared to other methods, making its module cohesion and module diversity distributions less comparable. Box-and-whisker plots show the median (center line), interquartile range (box), and 1.5× interquartile range (whiskers).

cluster, as do oligodendrocyte markers *Trf*, *Mobp*, *Apod*, and *Mal* (module 4, Fig. 2B). The network also captured subtle intra-module differences of genes—within module 1, *Meg3* correlated more strongly to *Snap25*, whereas *Nap1l5* is more similar to *Cartpt* (Fig. 2C). Further, genes that bridge two modules together, such as *Itm2b* or *Sparcl1*, are evident and retrievable from the network.

Next, we compare the gene modules to cell type subclasses identified in the adult mouse cerebellum region of the Allen Brain Cell (ABC) Atlas[10] (Supplementary Fig. 3). Plotting the 10 Smoothie modules revealed distinct gene patterns in the mouse cerebellum (Fig. 2D, E), many of which correspond to a specific cell subclasses from the ABC Atlas[10]. Modules 2, 3, 4, and 5 correspond to granule cells, Purkinje cells, oligodendrocytes, and choroid plexus cells, respectively. Module 6 is expressed primarily in neurons outside the cerebellum; module 7 is widely expressed across all cell types; and module 8 contains markers of non-neuronal cell types, particularly Bergmann glial cells, astrocytes, and ependymal cells. Modules 9 and 10 are smaller, containing only 3 and 2 genes, respectively; module 9 appears to align with glutamatergic neurons and module 10 with monocytes. We illustrate how Gaussian smoothing impacts module detection with the example of two known Purkinje cell marker genes *Car8* and *Calb1*; smoothing enhances their Pearson correlation coefficient significantly (Fig. 2F).

We benchmarked Smoothie against three leading methods for identifying gene modules from spatial transcriptomics data—Hotspot, STAMP, and hdWGCNA[4,11,12]. Mirroring previous gene module benchmarking approaches, we measure: (i) module cohesion (gene similarity within each

module), (ii) module diversity (how distinct each module is from the others), (iii) module count, and (iv) total gene recall in the clustering analysis (Methods)[11,12]. Smoothie identifies more gene modules with ~1.5x higher gene recall than Hotspot and STAMP, while achieving higher module cohesion and comparable module diversity (Fig. 2G). hdWGCNA assigns much fewer genes into modules—about 10x less gene recall than Smoothie and 5x less gene recall than Hotspot—making module cohesion and diversity measurements less comparable against the other three methods (Fig. 2G, Supplementary Fig. 4). We further compared Smoothie to Hotspot in terms of module agreement, gene similarity rankings, and runtime efficiency. Hotspot detected 6 gene modules in the cerebellum dataset, that align with Smoothie's modules 1–6 but did not recover modules 7–10 (Supplementary Fig. 3). To examine pairwise gene similarity ranks, we focused on 3 genes (*Car8, Malat1*, and *Ttr*) and compared the top 300 most similar genes identified by both Smoothie and Hotspot (Supplementary Fig. 5). While most top-ranked genes were shared by both methods, Smoothie identified several sparse genes (<500 UMIs) whose expression patterns visibly correlated with each gene of interest but which Hotspot disregarded due to low autocorrelation scores. Conversely, Hotspot sometimes incorrectly prioritized genes whose expression diverged from the gene of interest. Finally, runtime tests varying the number of genes and spatial locations confirmed Smoothie was significantly faster than Hotspot across all tested inputs, with over 100-fold runtime improvements for the larger tested inputs (Supplementary Fig. 6).

Smoothie finds diverse gene modules and new gene associations in the E16.5 mouse embryo

Next, we applied Smoothie to the larger and more complex 0.5 μm-resolution dataset of a E16.5 mouse embryo, which comprises 175,750,455 DNA nanoballs (≥ 1 UMIs) and 20,203 genes (≥ 100 UMIs tissue wide)[13]. While most analyses of such large data require binning or cell-segmentation preprocessing, Smoothie's memory and runtime efficiency enabled direct smoothing over all DNA nanoballs in ~1 hour on a standard CPU using ten parallel processes (Supplementary Fig. 6). After smoothing with a 20 μm Gaussian standard deviation, Smoothie identified 430,449 spatial gene correlations among 5944 genes using a PCC threshold of 0.4 (Supplementary Data 3). Of those genes, 3487 had a correlation above a PCC threshold of 0.5, which we visualize in the network (Fig. 3A, B). Random walk network clustering on the PCC > 0.4 network with soft power $\beta = 2$ led to 272 gene modules, with 89 modules having 3+ genes and 62 modules having 5+ genes (Fig. 3B). This number far exceeds the 35 modules reported in the original publication. We visualized module scores for the 16 largest spatial gene modules (Fig. 3C) and generated a gene pattern atlas for all 89 gene modules with 3 or more genes (Methods, Supplementary Fig. 8, Supplementary Data 4). These modules correspond to specific organs, sub-organ regions, cell types, or other gene programs. For example, gene module 1 corresponds to nerve tissue (*Map2, Gap43, Gad1* markers), module 4 to muscle tissue (*Myod1, Myf5, Myog, Pax7* markers), module 5 to liver tissue (*Alb, Afp* markers), and module 6 to outer epidermal tissue (*Krt1, Krt10, Lor* markers).

To evaluate clustering performance on a larger, more complex dataset, we again benchmarked Smoothie against Hotspot, STAMP, and hdWGCNA[4,11,12]. As before, we measure: (i) module cohesion, (ii) module diversity, (iii) module count, and (iv) gene count (Methods). Using recommended parameter settings from each method, we find that Smoothie identifies many more modules than competing methods, while achieving higher or comparable cohesion and diversity scores (Fig. 3D). Compared to Hotspot and STAMP, Smoothie identifies about two times more modules (of 3 or more genes) with higher module cohesion and diversity. Compared to hdWGCNA, Smoothie has over four times more modules with similar module cohesion and diversity (Fig. 3D). Visual inspection of Smoothie's gene pattern atlas reveals unique spatial patterns among the 89 modules of 3 or more genes (Supplementary Fig. 8). Further, gene set enrichment analysis and genome organization observations validate meaningful connections of the genes within modules. We also tested the other three methods across different hyperparameter settings to promote greater module identification levels comparable to Smoothie's module identification levels (Methods). We find that Hotspot and STAMP were able to generate module counts comparable to Smoothie; however, their module diversity values were lower, indicating many of their new modules had high pattern overlap with other modules (Supplementary Fig. 7). hdWGCNA did not achieve module counts comparable to Smoothie across tested hyperparameters. (Supplementary Fig. 7).

As Smoothie identified more gene modules compared to the original Stereo-seq paper and other benchmarked methods, we next looked for interesting biology in the dataset. First, we observed that Smoothie delineated sub-organ cell type patterns in the liver, revealing marker genes for hepatoblasts, neutrophils, megakaryocytes, and hematopoietic stem cells in modules 5, 35, 63, and 68, respectively (Fig. 3F). Similarly, in the kidney, Smoothie identified marker genes for tubule cells, collecting duct cells, podocytes, and kidney epithelial cells in clusters 21, 27, 64, and 71, respectively (Fig. 3F). Many of these sub-organ cell type patterns were uniquely identified by Smoothie compared to existing approaches. Examining module 14 in detail, we observed several astrocyte and radial glial cell marker genes in the brain's ventricle zones at an early stage in brain astrocytogenesis, prior to astrocyte migration and subtype differentiation[14–16](Fig. 3G). Module 14 contained 49 genes total, with a majority of such genes being new potential astrocyte markers during early phase astrocytogenesis (Supplementary Data 4). By identifying more spatial modules with higher gene recall, the Smoothie gene pattern atlas becomes a useful and extensive catalog of genes based on their spatial expression. This resource can reveal novel gene programs, new marker genes, and other biologically meaningful insights.

We then asked how many of the measured gene correlations represented novel associations. To explore this, we focused on the 22,281 hypothetical or predicted (unannotated) mouse genes in the *NCBI Gene* database lacking geneRIF functional annotation (as of Nov. 1st 2024). From this set, 1378 unannotated genes had ≥100 UMIs in the embryo dataset. Using a PCC > 0.4 cutoff, we detected 5914 pairwise correlations involving these unannotated genes, allowing us to assign modules and infer spatial annotations for 87 of them based on known tissue regions, cell types, or gene pathways (Fig. 3E). For example, we linked *Gm9947, Gm15533, Gm34081, Gm14635*, and *Gm14280* to muscle tissue (module 4), *Gm45941, Gm13629, Gm13425*, and *Gm12298* to *Sncg* neurons (module 10), *Gm26973* and *Gm19990* to lung tissue (module 15), *Gm7247* and *Gm14226* to the adrenal gland (module 16), and *Gm38505* and *Gm13889* to the subpallium brain region (module 29). We assigned modules to another 231 unannotated genes whose maximum PCC value fell between 0.2 and 0.4 by labeling each gene according to its highest-correlated neighbor in the PCC > 0.4 network. Overall, unannotated genes displayed diverse spatial patterns, appearing in 18 of the 20 largest modules and in 45 of 272 modules in total. If we also include genes with a maximum PCC of 0.2–0.4, the coverage extends to 72 modules. We provide a complete list of the unannotated gene correlations in Supplementary Data 5 with the corresponding module assignments in Supplementary Data 6.

Finally, we investigated the relationship between spatially correlated genes and genomic position, as genome organization and gene co-expression can be interconnected[17–20]. Using *UCSC Genome Browser*, first, we observed that spatial module 56 comprises 5 non-coding genes—*4930555F03Rik, Gm45321, Gm45323, Gm45341, C130073E24Rik*— that all fall within 2Mbp region on chromosome 8 (GENCODE *VM35*, Fig. 3H)[21,22]. Genes *4930555F03Rik* and *Gm45321* fall on the positive strand, while the other three are negative strand genes. Moreover, there is a > 300 kb gap between *4930555F03Rik* and *Gm45321*, and a > 200 kb gap between *Gm45341* and *C130073E24Rik*, suggesting that co-expression in this cluster may be driven by local regulatory mechanisms. In contrast, genes in other spatial modules are not co-located; for instance, module 64, which contains multiple kidney podocyte markers, includes four genes each situated on a different chromosome. We also examined the genomic locations of 59 gene pairs that formed "isolated" size-2 spatial modules in the network (i.e., each

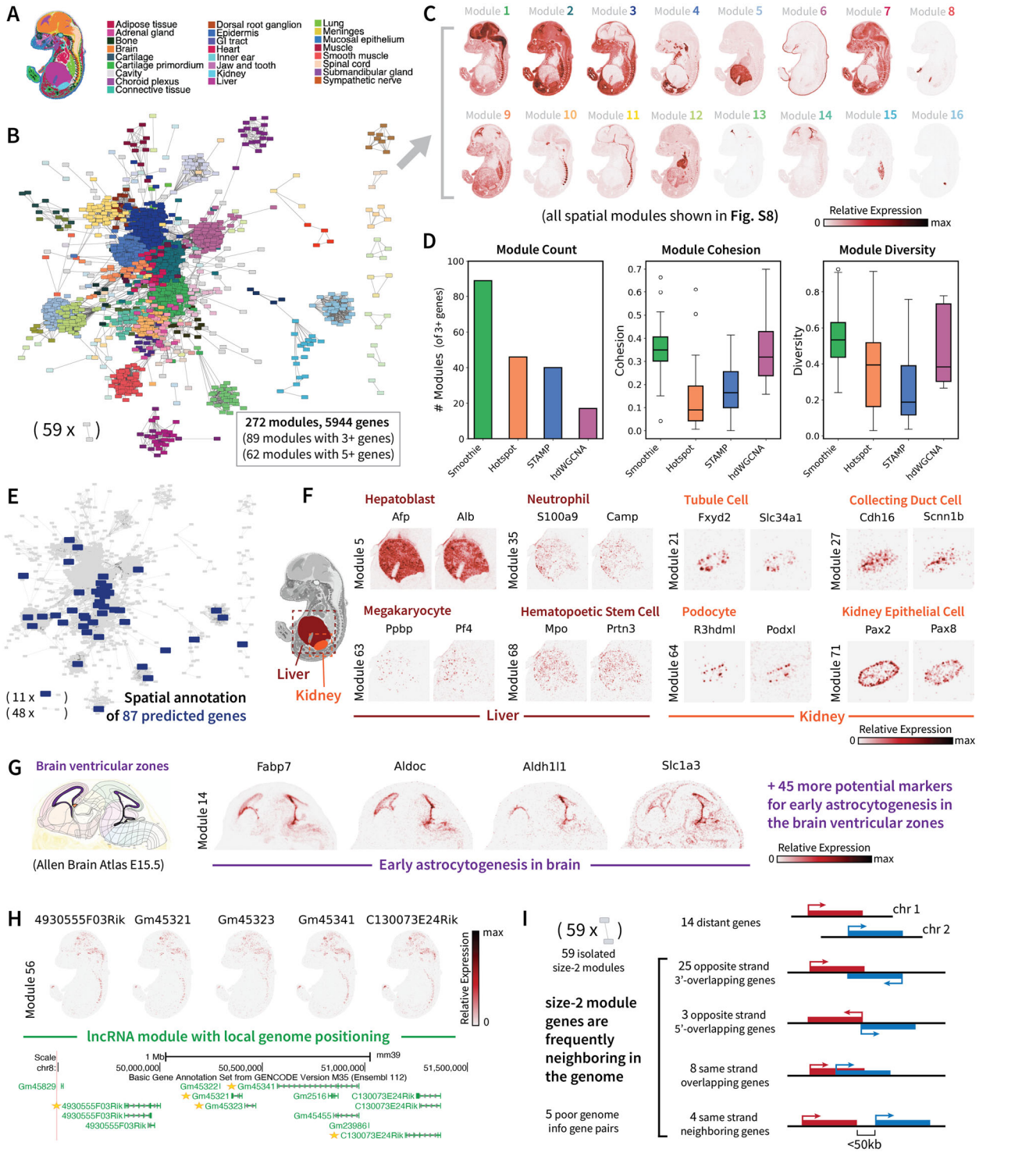

Fig. 3 | Smoothie reveals hundreds of spatial patterns in the developing E16.5 mouse embryo. A Region annotations from the original Stereo-seq paper. **B** Using a Pearson correlation coefficient (PCC) cutoff of 0.4 and soft power of 2, random walk network clustering finds 272 modules among 5944 genes, with 89 modules having 3+ genes and 62 modules having 5+ genes. Genes visualized in the network have at least one PCC value above 0.5 and are colored by their assigned module. **C** Module scores from the 16 largest modules. The module score reflects average smoothed expression of all genes in the module after rescaling each gene's maximum value to 1 (Methods). **D** Clustering benchmarks between Smoothie, Hotspot, STAMP, and hdWGCNA reveal that Smoothie identifies more modules than other methods, while achieving higher or similar cohesion and diversity levels. Total genes included per method: Smoothie, 5944 genes; Hotspot, 5000 genes; STAMP, 5500 genes; hdWGCNA, 3748 genes. Box-and-whisker plots show the median (center line), interquartile range (box), 1.5× interquartile range (whiskers), and outliers (points). **E** Smoothie spatially annotates 87 predicted or hypothetical mouse genes with no known gene function as of November 2024. **F** Examples of sub-organ level classification of spatial gene modules in the liver (left) and the kidney (right). **G** Several known astrocyte marker genes and new potential astrocyte marker genes reside in the brain ventricular zone in early astrocytogenesis. **H** Smoothie identifies a module of 5 lncRNAs that all reside within a 2 Mb window of chromosome 8. **I** Most size-2 gene modules exhibit near or overlapping genome positioning.

pair had a PCC > 0.5, and neither gene in a pair had a PCC > 0.4 with any other gene). Of these 59 pairs, 14 had genes situated on separate chromosomes or ≥15 Mb apart on the same chromosome, 44 comprised "overlapping" or closely spaced genes (<50 kb apart) on the same chromosome, and 1 pair was discarded because one member (Gm26776) lacked a known genomic locus. Closer inspection of the 44 pairs within 50 kb (using GENCODE *VM31* and *VM35* annotations) showed that 25 span opposite strands with 3′ overlap or proximity, 3 share opposite strands with 5′ overlap or proximity, 8 overlap on the same strand (potentially multi-mappers), 4 are consecutive neighbors on the same strand, and 4 pairs involved a chromosome accession number rather than a resolvable gene ID (Fig. 3I). This shows how genes located near each other in the genome are often enriched in these size-2 modules, with many pairs overlapping on opposite strands.

Smoothie characterizes the spatial patterns of the *Hox* family and *Slc* transporter family in the E16.5 embryo

We next used Smoothie to examine the spatial organization of gene sets of interest in the E16.5 mouse embryo, focusing first on the well-characterized *Hox* gene family. *Hox* genes are expressed during the development of almost all animals and exhibit defined anterior–posterior expression domains[23]. Out of the 39 mouse *Hox* genes, 33 showed non-random spatial correlations (along with 3 antisense transcripts), while the remaining 6 were too sparsely expressed to characterize. The *Hox* gene correlation network revealed that anterior *Hox* gene types 2–5 cluster on one side, whereas posterior *Hox* gene types 10–13 occupy the opposite side of the network (PCC > 0.2, Fig. 4B). Traversing the *Hox* network from left to right, we can visualize how the gene patterns change in their expression along the anterior-posterior axis, validating our spatial network approach (Fig. 4C).

We then analyzed the spatial distribution of solute carrier (*Slc*) membrane transport proteins (Fig. 4D). The *Slc* superfamily consists of over 400 genes across 65 families, each responsible for transporting specific ions, hormones, or other solutes[24–27]. Over 100 *Slc* genes have been implicated in human diseases, and their roles in molecular transport make them attractive targets for tissue-specific drug delivery[24–26]. Of the 356 *Slc* genes with 100+ UMIs in the embryo dataset, the PCC > 0.4 network organized 73 *Slc* genes into modules. Using a more permissive cutoff (PCC > 0.2), an additional 152 *Slc* genes were assigned to modules in this network. In total, 225 *Slc* genes were thus organized into 38 modules (Fig. 4D). These modules highlight *Slc* gene groups predominantly expressed in specific embryonic regions, such as the CNS (52 genes in module 1), liver (24 genes in module 5), meninges (23 genes in module 11), epidermis (10 genes in module 6), GI tract (9 genes in module 8), choroid plexus (7 genes in module 13), kidney tubule cells (6 genes in module 21), osteoblasts (3 genes in module 26), and others (Fig. 4E; Supplementary Data 7). Notably, many of these tissues are associated with fluid filtration—liver (blood), meninges (blood-brain barrier), choroid plexus (cerebrospinal fluid), GI tract (bile), and kidney tubule cells (blood). Other *Slc* genes show broad expression across multiple organs (e.g., modules 2, 3, and 7), potentially indicating housekeeping functions or expression in widespread cell types. This analysis aligns with a recent finding that no single *Slc* family is restricted to one tissue region[24]. Nevertheless, we detect enrichment of *Slc6, Slc13, Slc16*, and *Slc22* family genes in the meninges (module 11), whereas certain *Slc* families (e.g., *Slc25, Slc38*) exhibit more pervasive expression patterns.

Smoothie presents a way to analyze gene modules across multiple spatial experiments

Last, we show how Smoothie can integrate spatial transcriptomes from multiple samples to (i) find gene modules spanning multiple conditions, and (ii) quantify gene pattern stability or variability across conditions. To measure multi-dataset gene modules, we concatenate the normalized, smoothed gene surfaces of individual datasets into a combined dataset prior to pairwise gene correlation calculation (Methods). Applying this approach to the MOSTA mouse embryogenesis data at E13.5, E14.5, E15.5, and E16.5 revealed 782,486 correlated gene pairs (PCC > 0.4) among 6419 genes. Network clustering with soft power $\beta = 4$ identified 298 gene modules (111 modules with ≥ 3 genes; 69 modules with ≥ 5 genes; Fig. 5B; Supplementary Data 8, Supplementary Data 9)[13]. We used Enrichr to annotate modules by cell-type enrichment in several popular scRNA-seq atlases (Fig. 5C)[28–35]. Some examples include module 4 (skeletal muscle, 274 genes), module 14 (chondrocytes, 44 genes), module 25 (lung and intestine smooth muscle, 25 genes, module 37 (Type II pneumocyte, 12 genes), module 65 (all 5 *Zic* family genes), and module 92 (mesothelial cell layer, 3 genes), with all 111 modules (size ≥ 3) visualized in Supplementary Fig. 9.

To explore variation in gene expression across these four embryonic stages, we first generated pairwise gene correlation matrices for each dataset independently, capturing 7438 genes with at least one significant correlation (PCC > 0.4) in any dataset. We then quantified second-order correlations to ask: "Does gene *A* maintain similar correlation profiles among other genes in dataset D_1 and dataset D_2?" (Methods). This second order correlation quantifies gene *A*'s pattern stability across datasets (Fig. 5D). As shown in Fig. 5F, most genes had high stability (median values 0.7–0.9) across dataset pairs. For instance, the *Alb* liver marker remained consistently expressed, whereas *Mkrn1, Lor*, and *Gmza* varied significantly—*Mkrn1* expanded into arteries and heart by E16.5, *Lor* activated only at E16.5 in epidermal tissue, and *Gmza* disappeared at later stages because those tissue slices lacked the thymus (Fig. 5E). Combining the multi-dataset correlation network and the gene stability scores allowed us to identify temporally variable gene modules (Fig. 5G). For example, module 19 included numerous erythroid markers (e.g., *Mkrn1, Bpgm, Epb41, Adipor1*) that shifted from liver to blood vessel expression between E13.5 and E16.5, perhaps reflecting changes in erythroid cell abundance and location during embryogenesis[36] (Fig. 5H). Module 8 (epidermis) revealed temporal waves of gene activation—76 genes remained stable from E13.5–E16.5, 32 became activated from E14.5 onward, and 28 appeared only by E16.5 (including *Lor, Sbsn, Lipk*; Fig. 5I, Supplementary Data 10), revealing insights on the periderm's maturation from the E16.5 dataset[37,38].

Finally, we applied Smoothie to Slide-seq V2 spatial transcriptomics data of the mouse ovary collected at 8 distinct time point in the scope of hormone-induced ovulation (0–12 h post induction)[39,40]. Because each sample's follicle arrangement and stage in the ovary differ, integrating the data via coordinate warping or optimal transport alignment is both impractical and error-prone in this case. Smoothie identified 131,284 correlations (PCC > 0.45) among 2,689 genes, yielding 70 modules (33 modules with ≥5 genes; Supplementary Fig. 10, Supplementary Data 11, Supplementary Data 12). Several Smoothie modules agreed with the spatial cell type labels in the original paper[40] (Fig. 6A). For instance, module 2 (523 genes) corresponded to the cumulus oocyte complex (COC), capturing known markers (*Areg, Btc, Bmp15, Dazl, Ereg, Gdf9*) as well as new candidate regulators, including 15 *Fbxw* and 3 *Fbxo* genes implicated in SCF ubiquitin ligase function[41]. It has been shown that *Fbxw7* and *Fbxw24* knockout mice exhibit severe fertility defects[42,43]; however, the overall connection of the *Fbxw* and *Fbxo* gene families to the COC during oogenesis is understudied.

Further, we observed large temporal changes in gene expression within the antral mural granulosa cells (MGC) annotated in the original paper[40]. For example, Module 7 contained 173 genes with shared expression at 0–1 h, Module 1 had 612 genes with shared expression at 4 h, Module 10 included 60 genes expressed at 6–8 h, and Module 8 had 73 genes expressed between 6–12 h (Fig. 6B). These modules reveal wave-like changes in antral MGC gene programs over the 12-hour ovulation process, involving numerous regulatory factors and reported ovulation genes—such as *Pgr* and *Ptgs2* at 4 hours and *Adamts1* and *Edn2* between 6–12 h[44–47]. Plotting the average gene pattern stabilities for each of these modules reveals the antral MGC expression shifts too (Fig. 6C). Collectively, these examples illustrate Smoothie's utility for identifying, quantifying, and visualizing gene expression variation across diverse tissues.

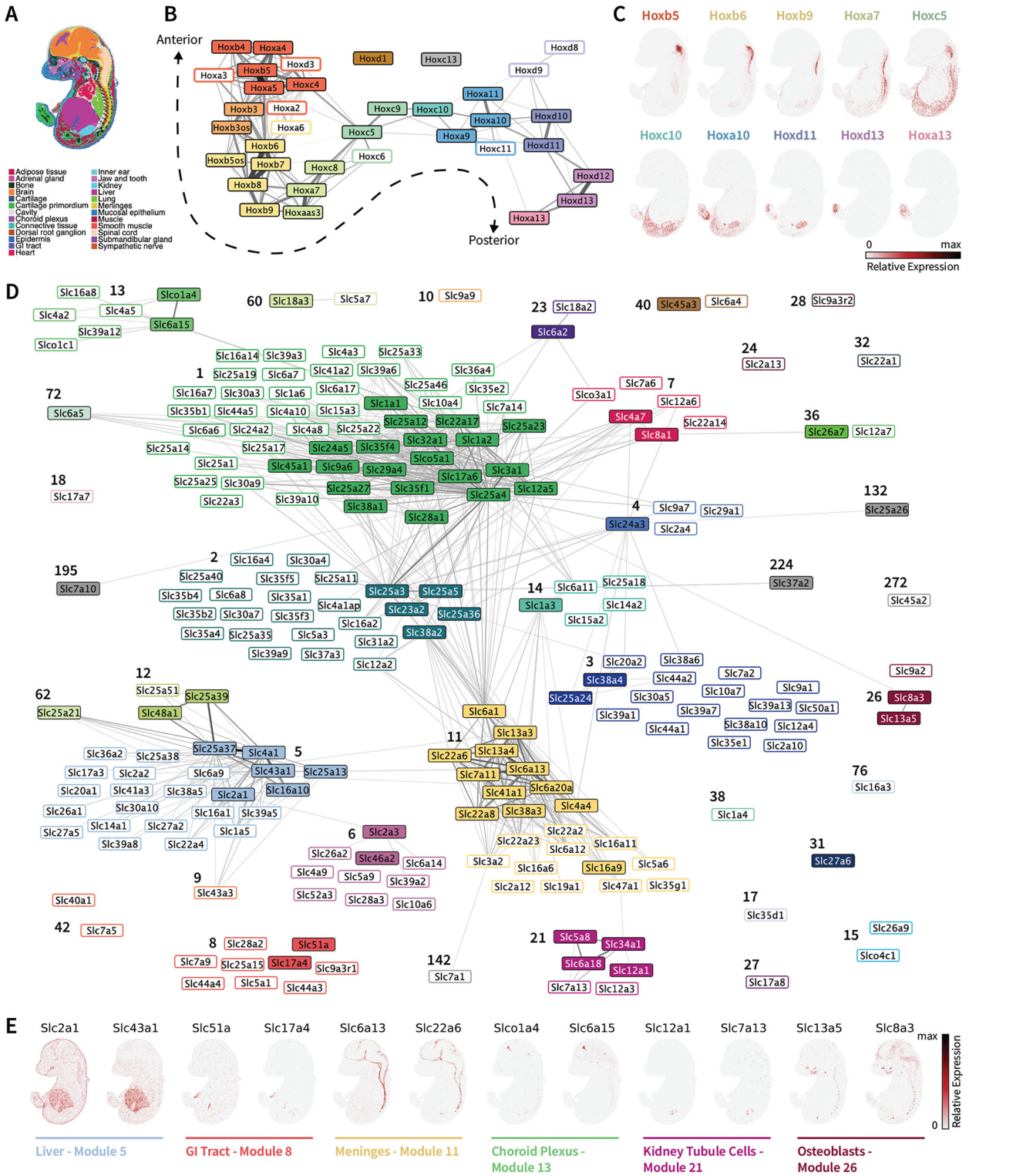

Fig. 4 | Smoothie comprehensively summarizes the spatial pattern distribution of the *Hox* gene family and the solute carrier (*Slc*) gene family in the E16.5 mouse embryo. A Region annotations from the original Stereo-seq paper. **B** The spatial *Hox* gene correlation network reveals the known anterior to posterior axis of expression. Edge width and darkness reflect the Pearson correlation coefficient (PCC); genes are colored by module; solid-colored genes have at least 1 correlation with PCC > 0.4; and hollow genes have no correlation with PCC > 0.4 but at least 1 correlation with PCC > 0.2. **C** Example *Hox* genes moving across the network from the anterior to posterior side. Plots show relative gene expression after normalization and smoothing. **D** The spatial *Slc* gene family correlation network reveals 38 modules of *Slc* genes. Edge and node specifications are identical to those in Fig. 4A. **E** Example *Slc* genes across different spatial modules. Plots show relative gene expression after normalization and smoothing.

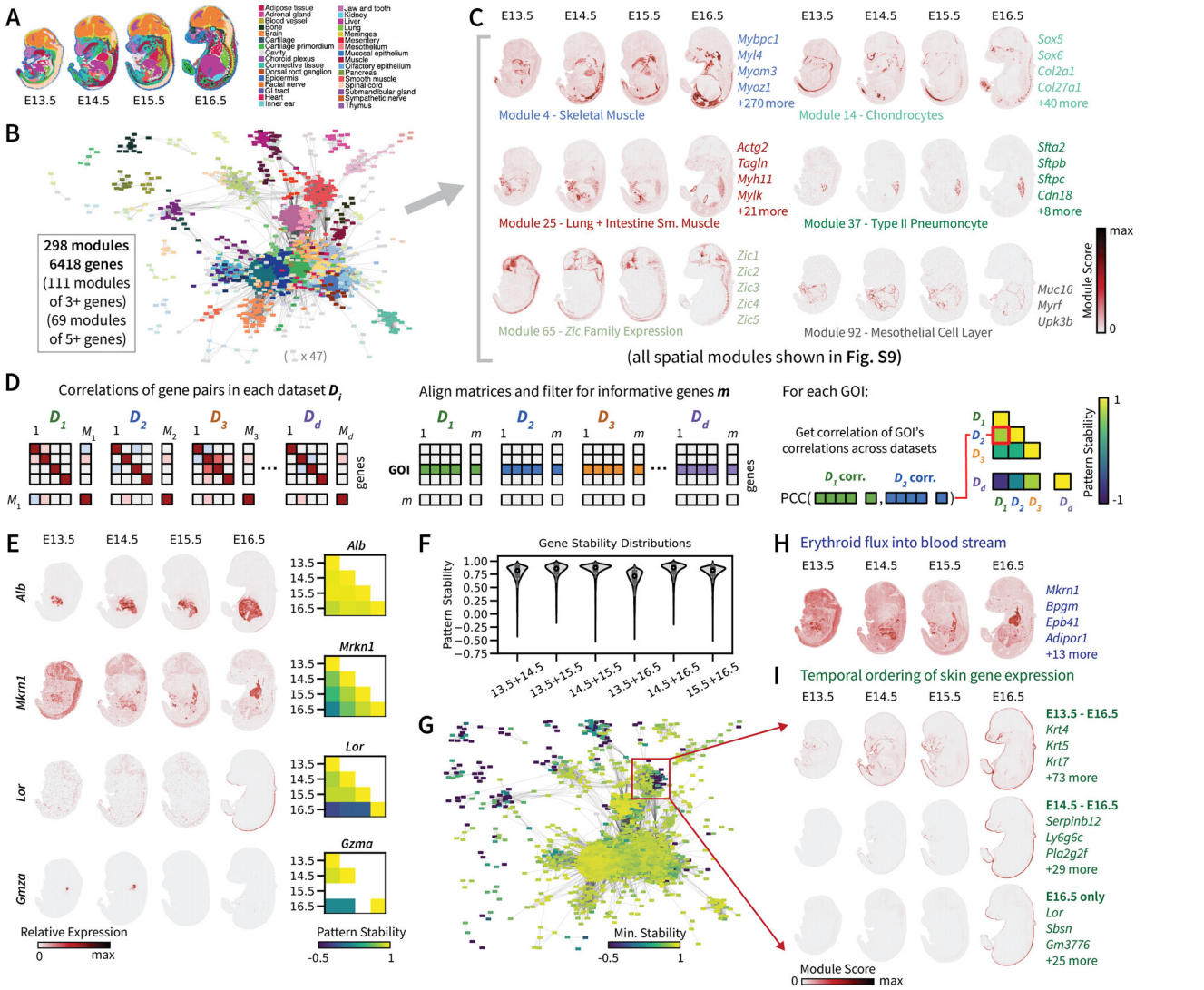

Fig. 5 | Smoothie finds multi dataset gene modules and ranks gene patterns from stable to dynamic. A Region annotations from the original Stereo-seq paper for the E13.5, E14.5, E15.5, and E16.5 embryos. **B** The gene correlation network created by comparing the joint gene patterns of all E13.5-E16.5 mouse embryos. **C** Example module plots from the network in 5B. Plots show the average 0 to 1 scaled genes for the module. **D** Overview of the second-order correlation logic used to rank gene patterns at stable or dynamic across spatial datasets of mismatched (x, y) coordinates. In short, intra-dataset pairwise gene correlations are used as gene embeddings that can be compared across datasets. **E** Examples of a stable gene pattern—liver marker *Alb*—and three dynamic gene patterns—*Mkrn1, Lor*, and *Gmza*. **F** The distributions of gene stability values across each dataset pair. Genes that are missing in one or the other dataset are not included in these distributions. **G** The spatial correlation network from 5A with gene stability node coloring reveals modules of dynamic gene patterns. Missing genes take the value of -1 for their stability in this network. **H** Example of a dynamic gene module containing 17 genes with several erythroid cell marker genes. **I** Gene stability analyses find temporal skin expression patterns. Some genes exhibit stable skin expression from E13.5-E16.5; some genes show skin expression from E14.5-E16.5; and some genes are only expressed at E16.5.

Discussion

In this work, we introduced Smoothie, a computationally efficient pipeline for measuring spatial gene correlation networks from denoised spatial transcriptomics data. Gaussian smoothing addresses data sparsity and noise, which expands the pool of analyzable genes in downstream correlation analyses. The ability to measure spatial correlations and assign gene modules to a larger number of genes (up to 5000 to 10,000 genes for some datasets) increases the potential for finding new biology. This sets Smoothie apart as an effective biological discovery tool, even for noisy, low-UMI datasets. We also show in benchmarking experiments that Smoothie achieves higher recall of diverse spatial gene modules compared to existing methods Hotspot, STAMP and hdWGCNA, finding 2x more modules in the Stereo-seq E16.5 developing mouse embryo. Further, while on bioRxiv, another independent research group reported that "*Smoothie is the most suitable co-expression matrix construction method for the MERFISH spatial transcriptomics data*" compared against methods SpaceX, WGCNA, and hdWGCNA[48].

We show that Smoothie's gene modules capture many facets of tissue biology. For instance, many spatial modules map onto single or closely related cell types (e.g., granule vs. Purkinje neurons in the cerebellum, or podocyte vs. tubule cells in the kidney), which can reveal new potential marker genes in those modules. The modules may also characterize understudied genes, which could be particularly impactful for studies of non-model organisms, where gene and functional annotations are often sparse. Even in the well-annotated mouse model and using data from just one experiment, we were able to spatially annotate hundreds of unknown genes. Interestingly, we also observed modules whose genes cluster by genome position. Integrating Smoothie's networks with motif analysis or Hi-C chromatin interaction data could further illuminate how local genome organization influences spatial transcript patterns and gene regulation.

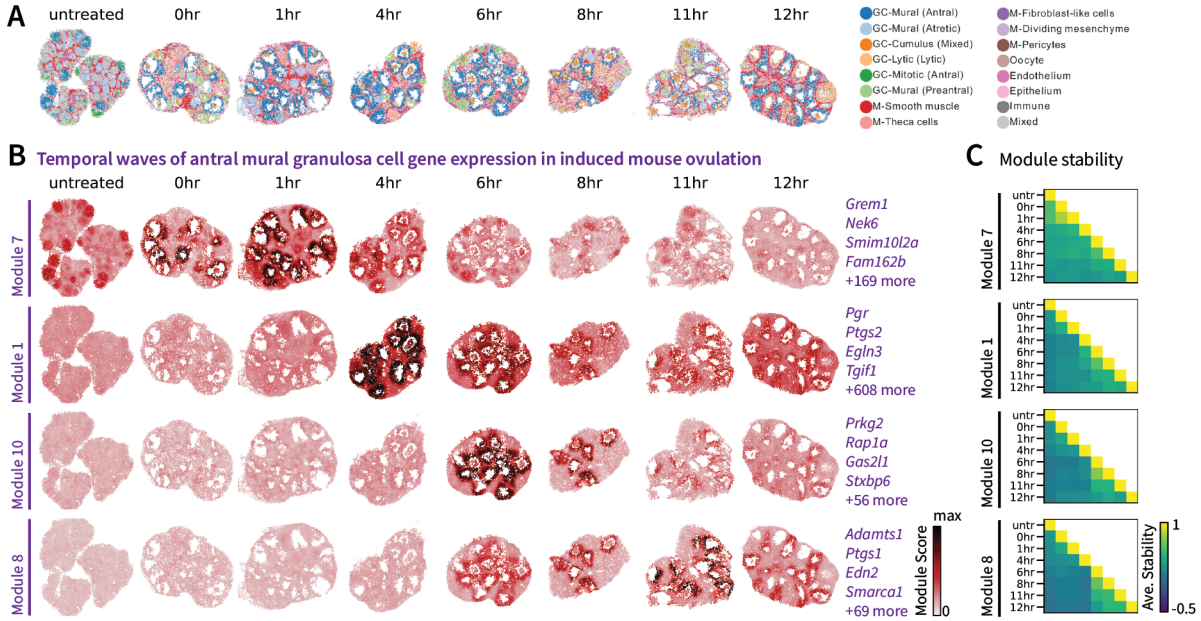

Fig. 6 | Smoothie finds dynamic gene programs within the antral mural granulosa cells (MGC) during induced mouse ovulation. A Cell type annotations from the original paper. **B** Smoothie identifies temporally restricted waves of gene expression in the antral mural granulosa cells throughout 0–12 h induced ovulation. The module score 0-max scale is shared across datasets. **C** Average gene pattern stabilities across datasets reveal the antral MGC expression shifts for each of the modules in panel 6B.

Finally, Smoothie also enables targeted investigations of particular genes, families, or pathways. For example, our analyses of the *Hox* and *Slc* gene families highlight the possibility to zoom in on a gene subset of interest, identifying submodules within the correlation network. Gene set analyses can combine a targeted subset network (for clarity) with the broader correlation network (for context), both of which Smoothie supports natively.

Another major application of Smoothie is multi-sample analysis and data integration. While numerous solutions have been proposed for integration of single-cell transcriptomic data, spatial transcriptomics has few robust solutions to integrate data across replicates, time points, conditions, or species[12,49]. Morphing of multiple tissues to match spatial patterns has been proposed, but this is error prone. Single-cell segmentation followed by integration is another possibility, but this is sensitive to any imperfections in segmentation, which can be substantial. Here, we show that the abstraction of the raw data enabled by measuring gene correlation networks provides a compelling avenue for data integration. We propose and implement two methods: (i) measuring gene correlations of all smoothed datasets together to find multi-dataset gene modules, and (ii) measuring second-order gene correlations across datasets to compare gene patterns across datasets. Such approaches to data integration can enable efficient comparison of spatially resolved transcriptomes from multiple tissues and time points. In addition, cross-species integration of spatial gene network data is an avenue for future investigation that can be impactful. Because measuring gene networks does not require consideration of the tissue architecture or biology, integration of spatial gene correlation networks may form the basis for the construction of large-scale spatial transcriptomic atlases of organ systems and organisms, akin to what has been accomplished with single-cell transcriptomic atlases.

We have shown that Smoothie is both effective and scalable on platforms ranging from Slide-seq (10 μm resolution) to Stereo-seq (0.5 μm resolution), providing confidence in its utility for other ST technologies—such as OpenST (0.6 μm), Visium-HD (2 μm), and beyond[9,13,39,50]. Further, we foresee that Smoothie's framework is adaptable to a range of other applications. Briefly, Smoothie could find correlations between spatial distributions of any data modality, e.g., protein expression, chromatin accessibility, metabolite levels, etc., and can thereby support multi-omics data integration. Furthermore, as the Gaussian smoothing kernel can be adjusted to account for three-dimensional distances, 3D spatial transcriptomic datasets could be analyzed via Smoothie. Last, Smoothie's ability to recover biological signals in noisy datasets will make it useful for applications where tissue quality can be poor, including clinical applications. As spatial omics technologies improve in resolution and scope, we anticipate Smoothie's utility will only grow, enabling adoption of a "forward spatial transcriptomics paradigm"—one in which data-driven, network-based approaches reveal novel biology rapidly across samples of different phenotypes or conditions.

Methods

Single dataset smoothie pipeline

Gaussian smoothing. Smoothie first performs Gaussian smoothing on each gene in the normalized spatial count matrix X to reduce data noise and sparsity. This process entails moving a 2D Gaussian distribution around the spatial data and taking a weighted average of the gene values under the distribution to generate the smoothed gene surface. The standard deviation σ of the 2D Gaussian specifies the degree of smoothing (Supplementary Fig. 1). The smoothing radius ρ is the range of points extending from the center of the Gaussian to include in the smoothing process. To avoid extensive and unnecessary compute power, the radius ρ is set to be 3σ in the analysis, which retains 99.7% of the volume under the Gaussian. There is an additional variable τ that specifies the minimum number of coordinates required within the Gaussian's radius for smoothing to occur. This prevents dividing by a small sum of values which can inflate smoothed gene values in some cases. The effect of variable τ is similar to the computer vision and deep learning concept of applying convolution layers without padding, slightly shrinking the dataset in edge regions of the tissue. Altogether, a Gaussian centered at a given spatial coordinate (x, y) generates the smoothed gene value $\bar{g}_{(x,y)}$ as follows:

$$\bar{g}_{(x,y)} = \begin{cases} \dfrac{\sum_{\{(x_i,y_i)\in B_\rho\}} \left[\exp\left(-\frac{(x_i-x)^2+(y_i-y)^2}{2\sigma^2}\right)\cdot g_{(x_i,y_i)}\right]}{\sum_{\{(x_i,y_i)\in B_\rho\}} \exp\left(-\frac{(x_i-x)^2+(y_i-y)^2}{2\sigma^2}\right)} & \text{if } \left|B_\rho\right| \geq \tau \\ 0 & \text{if } \left|B_\rho\right| < \tau \end{cases} \quad (1)$$

where, $g_{(x_i,y_i)}$ is the unsmoothed gene value at a given point (x_i, y_i) and B_ρ is the set of points that fall within the radius ρ from the point (x, y). The denominator contains the sum of the Gaussian values at each point within radius ρ to normalize for the non-uniform densities of spatial datapoints found in many spatial transcriptomic experiments. The accuracy of smoothing across genes is supported by plotting the total gene value sums for each gene before and after smoothing, revealing a near perfectly correlated line (Supplementary Fig. 1).

For scalability to very large datasets, Smoothie has two implementations of Gaussian smoothing – in-place and grid-based smoothing. In-place smoothing centers a Gaussian at every single spatial location in the data and generates a smoothed value for each spatial location. Grid-based smoothing instead imposes a hexagonal grid onto the original data with stride s distance between neighboring grid points and then smooths only at the imposed grid points. For binned or cell-segmented data, in-place smoothing is effective and standard. Grid-based smoothing is designed for smoothing directly on the higher-resolution platforms' data output, such as the 100 s of millions of Stereo-seq DNA nanoballs. In the process of smoothing, the sparse gene count matrix is converted into a dense matrix, so grid-based smoothing is required for matrices with several millions of spatial locations.

To summarize, the input parameters to Smoothie consist of the spatial gene count matrix X, Gaussian standard deviation σ, Gaussian radius ρ, Gaussian coordinate threshold τ, a Boolean value b to specify in-place or grid-based smoothing, and the hexagonal grid stride s for grid-based smoothing only. In practice, these parameter choices can be reduced to be in terms of just X and σ for simplicity and effective results:

$$\text{In}-\text{place smoothing}: \boldsymbol{X} = X, \boldsymbol{\sigma} = \sigma, \boldsymbol{\rho} = 3\sigma, \boldsymbol{\tau} \in [25, 100], \boldsymbol{b} = \mathit{False}, \boldsymbol{s} = \mathit{Null}$$

$$\text{Grid}-\text{based smoothing}: \boldsymbol{X} = X, \boldsymbol{\sigma} = \sigma, \boldsymbol{\rho} = 3\sigma, \boldsymbol{\tau} \in [25, 100], \boldsymbol{b} = \mathit{True}, \boldsymbol{s} \in [\sigma/2, \sigma]$$

Effective choices for σ could range from 10–40 μm, with 20–30 μm working well across tested platforms. Smoothie includes a function to test different smoothing degrees, performing the 2D Fourier transformation to measure and optimize gene expression "smoothness". The radial power spectrum output from a Fourier transformation measures the abundance of waves across frequencies that make up the image. Less high frequency, shorter wavelength waves indicate a smoother image. The radial power spectrum's spectral centroid, spectral slope, and fraction of high frequency waves (>0.025 cycles/μm) together can quantify the levels of noise or smoothness in the image (Supplementary Fig. 11). This approach finds 20-30 μm to be an optimal choice for σ across tested platforms. Further, comparing correlations of known spatially co-expressed cell type markers with correlations of random gene pairs across different levels of smoothing can be used to validate an optimal range for σ (Supplementary Fig. 1). The choice of τ depends on data sparsity following preprocessing and the choice of σ. A τ value of 25–100 is suitable for most datasets.

Gaussian smoothing scalable implementation. Smoothie generates the smoothed values for all genes in parallel with matrix multiplication. Further, generating smoothed gene values only requires information on the points within radius ρ from the center of the Gaussian, so different regions of the dataset can be smoothed in parallel. The runtime complexity for smoothing is $O(N' \cdot G \cdot \rho^2 \cdot \log(N)/P)$, where N is the number of spatial spots in the input, N' is the number of spatial spots in the smoothed output, G is the number of genes, and P is the number of parallel processes. Smoothie benefits from efficient KD-tree queries that introduce only a logarithmic dependence on N. For in-place smoothing, $N' = N$, and for grid-based smoothing $N' < \mathrm{N}$. Memory complexity of the smoothed output is $O(N' \cdot G)$, and a chunk count parameter C allows for efficient smoothing of smaller data segments, minimizing memory overhead during multi-processing. This combination of parallelization and chunking enables Smoothie to achieve fast, scalable runtimes while maintaining low memory requirements, even for large Stereo-seq datasets with hundreds of millions of DNA nanoballs.

Permutation tests to find significant gene correlations. With the gene correlation matrix R, Smoothie determines which spatial gene correlations are significant to include in downstream analyses. While each gene pair's Pearson correlation coefficient (PCC) is accompanied by a p value, the smoothing process artificially introduces a positive offset in correlation, preventing the p values from accurately defining a significance cutoff. Instead, Smoothie uses a permutation test to randomly shuffle the spatial coordinates of the input count matrix X, generating X_{shuffled}. Smoothie performs Gaussian smoothing with the same input parameters on both X and X_{shuffled} and generates pairwise gene correlation matrices R and R_{shuffled}. From the distribution of all randomized gene pair correlations in R_{shuffled}, the 99.9$^{\text{th}}$ percentile cutoff $PCC_{99.9}$ is used as a lower bound to select for significant, non-random gene correlations in the smoothed true dataset gene correlation distribution (Supplementary Fig. 1). There is also an option to shuffle cell-sized bins of spatial coordinates for subcellular resolution spatial platforms like Stereo-seq. Non-binned shuffling of subcellular coordinates would be a poor choice as there is co-expression of genes occurring within any given cell.

Creating and clustering the gene correlation network. Let X be a spatial count matrix with gene set $\mathcal{G} = \{g_1, g_2, \ldots g_G\}$; let R be the pairwise Pearson correlation coefficient matrix derived from a smoothed version of X; and let PCC^* be a hard threshold for the matrix R. Before clustering, each edge weight $R_{ij} > PCC^*$ between g_i and g_j is transformed by a soft power transformation $f_{PCC^*,\beta}$ to generate transformed edge weight R'_{ij}:

$$R'_{ij} = f_{PCC^*,\beta}\left(R_{ij}\right) = \left(\frac{R_{ij} - PCC^*}{1 - PCC^*}\right)^{\beta} \tag{2}$$

where $\beta \geq 1$ is the soft power. The transformation $f_{PCC^*,\beta}$ is monotonic and creates a sparse weighted network that emphasizes stronger correlations while suppressing weaker ones. After applying transformation $f_{PCC^*,\beta}$, the network used for clustering has node set V and edge set E:

$$V = \{g_1, g_2, \ldots g_G\}, E = \{(g_i, g_j, R'_{ij}) | R_{ij} > PCC^*\}$$

Gene modules are identified using the Infomap random network walk clustering algorithm implemented in the Python package *igraph*[7,51].

Smoothie includes a function to help the user select a choice of PCC^* and β. The function tests different combinations of PCC^* and β and measures: (i) gene recall, (ii) module count, (iii) network modularity, (iv) average gene margin, and (v) fraction of positive gene margin (Fig. S12). Network modularity Q measures how well dense gene correlations are captured within modules and is defined as:

$$Q = \sum_{c=1}^{n}\left[\frac{L_c}{m} - \gamma\left(\frac{k_c}{2m}\right)^2\right] \tag{3}$$

where L_c is the sum of intra-community edge weights for community c, k_c is the sum of weighted degrees for nodes within community c, m is the total edge weight between all nodes, n is the number of communities, and γ is a resolution parameter set to 1[52]. The gene margin, akin to the Silhouette score, for a gene g_i is defined as:

$$\text{Margin}_i = \frac{A_i - B_i}{\max(A_i, B_i)} \tag{4}$$

with A_i being the mean transformed correlation of gene g_i with other genes in its assigned module and B_i being the mean transformed correlation of gene g_i with genes in the next most similar module. Specifically,

$$A_i = \frac{1}{\left|\mathcal{G}_{M_i}\right| - 1}\sum_{j \in \mathcal{G}_{M_i}, j \neq i} R'_{ij} \tag{5}$$

$$B_i = \max_{k \neq M_i}\left(\frac{1}{|\mathcal{G}_k|}\sum_{j \in \mathcal{G}_k} R'_{ij}\right) \quad (6)$$

where M_i is the module assigned to gene g_i; $\mathcal{G}_k$ is the set of genes in module k; and R'_{ij} is the soft power transformed edge weight of gene g_i and gene g_j. The gene margin has range $[-1, 1]$, with more positive values indicating more confidence in the correct module assignment.

Selecting hard cutoff PCC^* involves balancing the user's goals of achieving higher gene recall (by selecting a lower PCC^*) versus achieving higher modularity and gene margin confidence (by selecting a higher PCC^*). Higher values of β favor more modular networks by making strong correlations more dominant in clustering.

Visualizing the correlation network. Smoothie's network output files can be imported into Cytoscape for visualization[53]. The "Prefuse Force Directed Layout" setting positions nodes in the network, and nodes may then be colored by their gene module assignment, chromosome number, binary values specifying gene set membership, or other variables.

Multiple dataset smoothie pipeline

Gaussian smoothing. Given a set of d spatial count matrices to integrate, the first step after normalization is to smooth each spatial count matrix in the set with equal smoothing parameters, as described above. Let $\{X_1, X_2, \ldots, X_d\}$ be the set of smoothed spatial count matrices, where each X_i has N_i spatial spots and G_i genes after smoothing.

Creating and clustering the gene correlation network for multiple datasets. The set of smoothed count matrices can be concatenated across spatial coordinates to create one smoothed count matrix X_{concat} with N_{concat} spatial spots (rows) and G_{concat} genes (columns). Here N_{concat} is the total number of points across all matrices and G_{concat} is the number of unique genes across all matrices. The creation of X_{concat} is such that gene columns are aligned across datasets during concatenation, and if a gene is missing in a given dataset, that dataset's portion of the gene column is filled with zeros. With X_{concat}, Smoothie can then create the pairwise gene correlation matrix R as before using a spatially permuted version of X_{concat} to determine the $PCC_{99.9}$ threshold of R_{shuffled}. After creating the weighted network W using correlation matrix R, hard threshold $PCC^* \geq PCC_{99.9}$, and soft power β, as explained before, random network walks clustering reveals spatial gene modules and the network can be visualized for interpretation.

Cross dataset gene pattern embedding comparisons. Here, we describe an approach to measure the similarity of two genes' spatial expression patterns across different spatial datasets. Given a set of smoothed spatial gene count matrices $\{X_1, X_2, \ldots, X_d\}$, where each X_i has N_i spatial spots and G_i genes, these datasets have different tissue morphologies and different (x, y) coordinates. Thus, the smoothed gene surfaces between two distinct count matrices cannot be directly compared via Pearson correlation. Instead, Smoothie leverages a second-order correlation approach. After generating a gene pair correlation matrix R_i for each smoothed count matrix ($R_1 \in \mathbb{R}^{G_1 \times G_1}$, $R_2 \in \mathbb{R}^{G_2 \times G_2}$, …, $R_d \in \mathbb{R}^{G_d \times G_d}$), Smoothie creates aligned versions of the correlation matrices for each dataset ($R^*_1 \in \mathbb{R}^{G^* \times G^*}$, $R^*_2 \in \mathbb{R}^{G^* \times G^*}$, …, $R^*_d \in \mathbb{R}^{G^* \times G^*}$) where (1) row j and column j are the same gene across all R^*_i and (2) G^* is the magnitude of the gene set $\mathcal{G}^* = \{g_1, g_2, \ldots g_{G^*}\}$ of genes that have a Pearson correlation coefficient above a threshold PCC^* in at least 1 dataset. Genes that are missing (< 100 UMIs) in one dataset but significant in another dataset have their missing gene's correlations filled with *Null* values. The threshold PCC^* is determined considering the $PCC_{99.9}$ values determined with permutation tests for each individual dataset.

Genes in $\mathcal{G}^*$ can then be compared across datasets by measuring the correlation of any two rows from the aligned matrices $R^*_1, R^*_2, \ldots, R^*_d$. In other words, for each dataset, a given gene g in $\mathcal{G}^*$ has a gene embedding row vector $\Phi_g = [e_1, e_2, \ldots, e_{G^*}]$ where e_i is the correlation value of gene g and gene g_i, and these embedding vectors are comparable across datasets. Prior to comparing embedding vectors, the Fisher's Z transformation is applied to all correlation values across all R^*_i to stabilize the variance of the correlation distributions. If two gene embeddings are similar across two datasets, then those genes have similar spatial expression patterns, whereas dissimilar embeddings indicate dynamic spatial expression patterns. The formula for measuring gene similarity between gene A in R^*_1 and gene B in R^*_2 is simply the Pearson correlation coefficient of the two gene embedding vectors $\Phi_{1A} = [e_{11}, e_{12}, \ldots, e_{1G^*}]$ and $\Phi_{2B} = [e_{21}, e_{22}, \ldots, e_{2G^*}]$, using only embedding features with finite values in both vectors in the calculation:

$$sim(\Phi_{1A}, \Phi_{2B}) = \frac{\sum(e_{1i} - \bar{e}_1)(e_{2i} - \bar{e}_2)}{\sqrt{\sum(e_{1i} - \bar{e}_1)^2 \sum(e_{2i} - \bar{e}_2)^2}} \quad (7)$$

This logic relies on two key assumptions:

1. Gene embeddings are comparable across datasets. Comparing genes across datasets using their intra-dataset gene correlations requires that a majority of genes have comparable or "stable" expression patterns across datasets (i.e., most gene embedding features are comparable). Each comparable embedding feature contributes to the success of the second order correlation approach, whereas unaligned embedding features add noise to the calculation. By using only the $\mathcal{G}^*$ significantly correlated genes for this analysis, Smoothie filters out genes with less structured and more noisy patterns, which increases the ratio of stable to unstable genes. In some cases, the corresponding embedding features for known unstable genes may be filtered out prior to analysis.
2. Gene embeddings adequately cover the complexity of gene patterns. Gene embeddings must also span the complexity of gene patterns present in the tissue slices with near equal pattern weights. Here, embeddings are designed as gene correlations with all other genes in $\mathcal{G}^*$, maximizing pattern diversity without having to add any synthetic features. Since some modules contain more genes than others, a parameter $e_{\max}$ limits the number of genes per module used as embedding features. If a module exceeds $e_{\max}$, its features are randomly down sampled. After down sampling, each matrix R^*_i retains the same number of genes to compare but has fewer embedding features for cross-dataset comparisons.

Method benchmarking

We benchmark Smoothie against leading spatial gene module detection methods Hotspot, STAMP, and hdWGCNA using both the Slide-seq mouse cerebellum dataset and the Stereo-seq E16.5 mouse embryo dataset. Hotspot identifies genes with significant local autocorrelation and then groups those genes into modules with a local correlation statistic. STAMP uses a deep generative topic model with a simplified graph convolution network to output spatially organized topics with associated gene modules. hdWGCNA groups spatial locations into metacells and then applies the WGCNA pipeline to find gene correlations and identify modules. We did not benchmark Smoothie against standalone denoising approaches, as Hotspot, STAMP, and hdWGCNA already include built-in mechanisms to handle data sparsity as part of their pipelines.

Using the gene module assignments from each method, we measured: (i) module cohesion—gene similarity within each module—(ii) module diversity—how distinct each module is from the others—(iii) module count, and (iv) total genes included in the modules. Let x_i be the gene expression vector across locations for gene g_i; let G_m be the set of genes in module m; and let R_{ij} be the Pearson correlation coefficient between the gene expression vectors for gene g_i and gene g_j.

Module score. Each gene expression vector is rescaled by its maximum value, and the module score S_m for a given module m is defined as the

average of the rescaled gene vectors:

$$S_m = \frac{1}{|G_m|}\sum_{i\in G_m}\frac{x_i}{\max(x_i)} \quad (8)$$

Module cohesion. Module cohesion measures the average intra-module edge weight:

$$\text{Cohesion}_m = \frac{1}{|G_m|\,(|G_m|-1)}\sum_{i,j\in G_m i\neq j} R_{ij} \quad (9)$$

Module diversity. Module diversity quantifies how distinct each module is from the others:

$$\text{Diversity}_m = 1 - \max_{n\neq m}\left(\text{PCC}\left(\text{S}_\text{m},\text{S}_\text{n}\right)\right) \quad (10)$$

where $\text{PCC}\left(\text{S}_\text{m},\text{S}_\text{n}\right)$ is the Pearson correlation coefficient of module score vector S_m with module score vector $S_n \neq S_m$. A higher diversity means the module is more spatially unique compared to all other modules.

The formulas for cohesion and diversity use pairwise gene correlations to assess the quality of each method's output modules. We calculate cohesion and diversity in two ways: (i) using the unsmoothed normalized data to calculate gene correlations and (ii) using the denoised normalized data (after 20 μm Gaussian std. dev. smoothing) to calculate gene correlations. Across all benchmarked methods, using the denoised data resulted in more positive module cohesion distributions and less positive module diversity distributions. This is because the distribution of pairwise gene correlations is broader in the denoised data compared to the unsmoothed data. Thus, there can be larger cohesion values—stronger correlated gene patterns—and smaller diversity values—stronger correlated module scores—since the magnitude of gene correlations is enhanced. Module cohesion and diversity calculations using (i) unsmoothed and (ii) denoised gene correlations are included in the Supplementary Figs., with the latter calculation method used in the main figures.

Software requirements

Smoothie was implemented using Python 3.10.12, along with the following packages: anndata (v0.9.2)[54], igraph (v0.10.6)[51], numpy (v1.23.5)[55], pandas (v1.5.3)[56], scanpy (v1.9.3)[57], and scipy (v1.9.3)[58], matplotlib (v3.7.1)[59]. The pipeline is compatible with newer versions of these packages too.

Data resources, preprocessing, and normalization

Mouse cerebellum data, Slide-seq. The Slide-seq Puck_180819_12 dataset .h5ad file was obtained from the Hotspot Github page https://hotspot.readthedocs.io/en/latest/Spatial_Tutorial.html. This data is also available from the original paper here https://singlecell.broadinstitute.org/single_cell/study/SCP354/slide-seq-study#study-summary[9]. Genes expressed in at least 50 different spatial locations were kept for the analysis. Data was normalized with counts per thousand (CPT) normalization and then log1p normalization.

Mouse embryo development data, Stereo-seq. The raw DNA nanoball datasets E13.5_E1S1_GEM_bin1.tsv.gz, E14.5_E1S1_GEM_bin1.tsv.gz, E15.5_E1S1_GEM_bin1.tsv.gz, and E16.5_E1S1_GEM_bin1.tsv.gz were obtained from https://db.cngb.org/stomics/mosta/download/ and converted to .h5ad files[13]. DNA nanoballs with 1 or more UMIs and genes with at least 100 total UMIs were kept. Data was normalized with only log1p normalization, as CPT normalization on DNA nanoballs with 1 to a few UMIs is nonsensical.

Mouse ovary timecourse data, Slide-seq V2. Sequencing data is available with GEO number GSE240271 and spatial barcode files are listed in https://github.com/madhavmantri/mouse_ovulation[40]. Spatial spots with at least 100 UMIs and genes with at least 50 total UMIs were kept. Data was normalized with CPT normalization and then log1p normalization.

Statistics and reproducibility

Slide-seq mouse cerebellum. In-place Gaussian smoothing was performed on Slide-seq beads with Gaussian standard deviation $\boldsymbol{\sigma} = 30$ μm, Gaussian radius $\boldsymbol{\rho} = 90$ μm, and minimum coordinates under Gaussian $\boldsymbol{\tau} = 100$. The PCC distribution of permuted spatial coordinate gene correlations found a $PCC_{99.9}$ cutoff of 0.2165. The correlation network was built with a $\boldsymbol{PCC}^* = 0.35$, and the clustering power $\boldsymbol{\beta} = 4$ was used for gene module identification.

Stereo-seq E16.5 mouse embryo. Grid-based Gaussian smoothing was performed on Stereo-seq DNA nanoballs with Gaussian standard deviation $\boldsymbol{\sigma} = 20$ μm, Gaussian radius $\boldsymbol{\rho} = 60$ μm, minimum coordinates under Gaussian $\boldsymbol{\tau} = 100$, and hexagonal grid stride $\boldsymbol{s} = 20$ μm. The PCC distribution of permuted spatial coordinate gene correlations found a $PCC_{99.9}$ cutoff of 0.1098. The correlation network was built with a $\boldsymbol{PCC}^* = 0.4$, and the clustering power $\boldsymbol{\beta} = 2$ was used for gene module identification.

Stereo-seq E13.5 to E16.5 mouse embryo time course. Grid-based Gaussian smoothing was performed on DNA nanoballs of each embryo dataset—E13.5, E14.5, E15.5, E16.5—with identical smoothing parameters of Gaussian standard deviation $\boldsymbol{\sigma} = 20$ μm, Gaussian radius $\boldsymbol{\rho} = 60$ μm, minimum coordinates under Gaussian $\boldsymbol{\tau} = 100$, and stride $\boldsymbol{s} = 20$ μm. For multi-dataset module detection, the $PCC_{99.9}$ cutoff generated from the combined permuted datasets was 0.406. The combined dataset correlation network was built with a $\boldsymbol{PCC}^* = 0.4$, and the clustering power $\boldsymbol{\beta} = 4$ was used for gene module identification. For gene embedding comparison analyses, the maximum $PCC_{99.9}$ cutoff generated among the permuted individual datasets was 0.116. The set of $\mathcal{G}^*$ genes for embedding comparison analysis were selected with cutoff $\boldsymbol{PCC}^* = 0.4$, and the maximum feature embeddings per module was capped at $\boldsymbol{e}_{\max} = 25$.

Slide-seq V2 mouse ovaries time course. In-place Gaussian smoothing was performed on the Slide-seq V2 beads of each ovary dataset—untreated, 0 h, 1 h, 4 hr, 6 h, 8 h, 11 h, and 12 h—with identical smoothing parameters of Gaussian standard deviation $\boldsymbol{\sigma} = 30$ μm, Gaussian radius $\boldsymbol{\rho} = 90$ μm, and minimum Gaussian coordinates $\boldsymbol{\tau} = 25$. For multi-dataset module detection, the $PCC_{99.9}$ cutoff generated from the combined permuted datasets was 0.420507. The combined dataset correlation network was built with a $\boldsymbol{PCC}^* = 0.45$, and the clustering power $\boldsymbol{\beta} = 4$ was used for gene module identification. For gene embedding comparison analyses, the maximum $PCC_{99.9}$ cutoff generated among the permuted individual datasets was 0.2075. The set of $\mathcal{G}^*$ genes for embedding comparison analysis were selected with cutoff $\boldsymbol{PCC}^* = 0.4$, and the maximum feature embeddings per module was capped at $\boldsymbol{e}_{\max} = 25$.

Benchmarking. For the Slide-seq adult mouse cerebellum benchmarks, we ran each method with default or optimized parameters. Specifically, we ran Hotspot with *min_gene_threshold* = 20 and *n_neighbors* = 300; we ran STAMP with *n_topics* = 6; and we ran hdWGCNA with *minModuleSize* = 10.

For the 20 μm binned E16.5 Stereo-seq embryo, we ran each benchmarked method 3 times—once with the default parameters and two additional times with slight parameter modifications to encourage greater clustering and more modules. Specifically, we ran Hotspot by selecting *min_gene_threshold* ∈ {20 (recommended), 10, 8} with n_neighbors=100; we ran STAMP with *n_topics* ∈ {40 (recommended), 80, 120}; and we ran hdWGCNA with *minModuleSize* ∈ {40 (recommended), 10, 5}.

STAMP outputs soft topic loadings for each spot and for each gene. For comparison to hard-partitioning methods—Smoothie, Hotspot,

hdWGCNA—we binarized each gene to the topic with the highest loading. Additionally, compared to other methods—Smoothie, Hotspot, hdWGCNA—STAMP does not perform feature selection of genes; rather, in the paper they examine only the top 20 genes from each topic. Here, we included the top k genes with the highest maximum value across topic loadings, where the choice of k is more comparable to the other methods.

For the benchmarks between only Smoothie and Hotspot, the Hotspot code was run as provided here https://hotspot.readthedocs.io/en/latest/Spatial_Tutorial.html. For runtime comparisons across different cell counts and gene counts, the same filtered count matrix was used for both Hotspot and Smoothie. Runtime experiments were run on a high-performance computing system with an AMD EPYC 7763 64-core processor and 1000 GB of RAM.

Reporting summary

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

Data availability

Supplementary data, including edge lists and node labels for all networks from this study, are available here https://doi.org/10.5281/zenodo.14933147[60].

Code availability

Smoothie is available as an open source package at https://github.com/caholdener01/Smoothie, including easy to use example scripts. The code for all analyses supporting this manuscript is available at https://doi.org/10.5281/zenodo.18305935[61].

Received: 15 May 2025; Accepted: 9 March 2026;
Published online: 26 March 2026

References

1. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 1 (2005).
2. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
3. Morabito, S. et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat. Genet* **53**, 1143–1155 (2021).
4. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E. & Swarup, V. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Rep. Methods* **3**, 6 (2023).
5. Van Dam, S., Võsa, U., Van Der Graaf, A., Franke, L. & De Magalhães, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* bbw139 https://doi.org/10.1093/bib/bbw139 (2017).
6. Mitchel, J. et al. Impact and correction of segmentation errors in spatial transcriptomics. *Nat. Genet.* **58**, 434–444 (2026).
7. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. USA* **105**, 1118–1123 (2008).
8. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).
9. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
10. Yao, Z. et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).
11. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446–456.e9 (2021).
12. Zhong, C., Ang, K. S. & Chen, J. Interpretable spatially aware dimension reduction of spatial transcriptomics with STAMP. *Nat. Methods* **21**, 2072–2083 (2024).
13. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792.e21 (2022).
14. Vivi, E. & Di Benedetto, B. Brain stars take the lead during critical periods of early postnatal brain development: relevance of astrocytes in health and mental disorders. *Mol. Psychiatry* **29**, 2821–2833 (2024).
15. Cahoy, J. D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
16. Akdemir, E. S., Huang, A. Y.-S. & Deneen, B. Astrocytogenesis: where, when, and how. *F1000Research* **9**, F1000 Faculty Rev (2020).
17. Ribeiro, D. M., Ziyani, C. & Delaneau, O. Shared regulation and functional relevance of local gene co-expression revealed by single cell analysis. *Commun. Biol.* **5**, 1–11 (2022).
18. Hurst, L. D., Pál, C. & Lercher, M. J. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet* **5**, 299–310 (2004).
19. Soler-Oliva, M. E., Guerrero-Martínez, J. A., Bachetti, V. & Reyes, J. C. Analysis of the relationship between coexpression domains and chromatin 3D organization. *PLOS Comput. Biol.* **13**, e1005708 (2017).
20. Ribeiro, D. M. et al. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.* **12**, 4842 (2021).
21. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
22. Mudge, J. M. et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
23. Philippidou, P. & Dasen, J. S. *Hox* Genes: choreographers in neural development, architects of circuit organization. *Neuron* **80**, 12–34 (2013).
24. César-Razquin, A. et al. A call for systematic research on solute carriers. *Cell* **162**, 478–487 (2015).
25. Aykac, A. & Sehirli, A. O. The role of the SLC transporters protein in the neurodegenerative disorders. *Clin. Psychopharmacol. Neurosci.* **18**, 174 (2020).
26. Zhang, Y., Zhang, Y., Sun, K., Meng, Z. & Chen, L. The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. *J. Mol. Cell Biol.* **11**, 1–13 (2019).
27. Morris, M. E., Rodriguez-Cruz, V. & Felmlee, M. A. SLC and ABC transporters: expression, localization, and species differences at the blood-brain and the blood-cerebrospinal fluid barriers. *AAPS J.* **19**, 1317 (2017).
28. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).
29. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
30. Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
31. Tabula Sapiens Consortium* et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
32. HuBMAP Consortium The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
33. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
34. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).
35. Hu, C. et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).

36. Yumine, A., Fraser, S. T. & Sugiyama, D. Regulation of the embryonic erythropoietic niche: a future perspective. *Blood Res.* **52**, 10–17 (2017).
37. Jacob, T. et al. Molecular and spatial landmarks of early mouse skin development. *Dev. Cell* **58**, 2140–2162.e5 (2023).
38. Nakazawa, S. et al. Suprabasin-null mice retain skin barrier function and show high contact hypersensitivity to nickel upon oral nickel loading. *Sci. Rep.* **10**, 14559 (2020).
39. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
40. Mantri, M., Zhang, H. H., Spanos, E., Ren, Y. A. & De Vlaminck, I. A spatiotemporal molecular atlas of the ovulating mouse ovary. *Proc. Natl. Acad. Sci. USA* **121**, e2317418121 (2024).
41. Kipreos, E. T. & Pagano, M. The F-box protein family. *Genome Biol.* **1217**, 111–122 (2000).
42. Zhao, H. et al. Deletion of Fbxw7 in oocytes causes follicle loss and premature ovarian insufficiency in mice. *J. Cell. Mol. Med.* **28**, e18487 (2024).
43. Wang, Y. et al. FBXW24 controls female meiotic prophase progression by regulating SYCP3 ubiquitination. *Clin. Transl. Med.* **12**, e891 (2022).
44. Park, C. J. et al. Progesterone receptor serves the ovary as a trigger of ovulation and a terminator of inflammation. *Cell Rep.* **31**, 107496 (2020).
45. Kim, J., Bagchi, I. C. & Bagchi, M. K. Control of ovulation in mice by progesterone receptor-regulated gene networks. *Mol. Hum. Reprod.* **15**, 821–828 (2009).
46. Robker, R. L. et al. Progesterone-regulated genes in the ovulation process: ADAMTS-1 and cathepsin L proteases. *Proc. Natl. Acad. Sci. USA* **97**, 4689–4694 (2000).
47. Cacioppo, J. A. et al. Granulosa cell endothelin-2 expression is fundamental for ovulatory follicle rupture. *Sci. Rep.* **7**, 817 (2017).
48. Liu, T. et al. Linking spatial omics to patient phenotypes at the population scale by BSNMani: Bayesian scalar-on-network regression with manifold learning. medRxiv preprint at https://doi.org/10.1101/2025.08.09.25333297 (2025).
49. Klein, D. et al. Mapping cells through time and space with Moscot. *Nature* 1065–1075 https://doi.org/10.1038/s41586-024-08453-2 (2025).
50. Schott, M. et al. Open-ST: High-resolution spatial transcriptomics in 3D. *Cell* **187**, 3953–3972.e26 (2024).
51. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).
52. Hagberg, A. A., Schult, D. A. & Swart, P. J. *Exploring Network Structure, Dynamics, and Function Using NetworkX* 11–15 (Pasadena, California, 2008). https://doi.org/10.25080/TCWV9851.
53. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).
54. Virshup, I. et al. anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).
55. Harris, C, R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
56. McKinney, W. *Data Structures for Statistical Computing in Python* 56–61 (Austin, Texas, 2010). https://doi.org/10.25080/Majora-92bf1922-00a.
57. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
58. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
59. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
60. Holdener, C. & De Vlaminck, I. Supplementary data for 'smoothie: efficient inference of spatial co-expression networks from denoised spatial transcriptomics data'. *Zenodo* https://doi.org/10.5281/ZENODO.14933147 (2025).
61. Holdener, C. & De Vlaminck, I. Analyses and scripts for 'smoothie: efficient inference of spatial co-expression networks from denoised spatial transcriptomics data'. *Zenodo* https://doi.org/10.5281/ZENODO.18305935 (2026).

Acknowledgements

C.H. and I.D.V. would like to thank Dr. Sumanta Basu for helpful input on statistical testing of Smoothie and Dr. Yi Ren for helpful input on the ovary time course dataset. C.H. would also like to thank Nicole Eikmeier and Jerod Weinman at Grinnell College for their network science and computer vision coursework that inspired the core idea of this paper. C.H. and I.D.V. would also like to thank all members of the De Vlaminck lab for helpful comments, discussion, and feedback. This work was supported by NIH R01AI176681, NIH R01AI176681, NIH R01AI189855 and CZI 2023-323354 to I.D.V. as well as NSF GRFP DGE – 2139899 to C.H.

Author contributions

C.H. and I.D.V. conceptualized the study. C.H. wrote the software. C.H. performed analyses with guidance from I.D.V.; C.H. and I.D.V. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-09898-z.

Correspondence and requests for materials should be addressed to Iwijn De Vlaminck.

Peer review information Communications Biology thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Dr. Adib Keikhosravi and Dr. Nilanjan Banerjee & Dr. Mengtan Xing. A peer review file is available.