Article

Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics

Lulu Shang [1,4], Peijun Wu [2,3,4] & Xiang Zhou [2,3] ✉

An essential task in spatial transcriptomics is identifying spatially variable genes (SVGs). Here, we present Celina, a statistical method for systematically detecting cell type-specific SVGs (ct-SVGs)—a subset of SVGs exhibiting distinct spatial expression patterns within specific cell types. Celina utilizes a spatially varying coefficient model to accurately capture each gene's spatial expression pattern in relation to the distribution of cell types across tissue locations, ensuring effective type I error control and high power. Celina proves powerful compared to existing methods in single-cell resolution spatial transcriptomics and stands as the only effective solution for spot-resolution spatial transcriptomics. Applied to five real datasets, Celina uncovers ct-SVGs associated with tumor progression and patient survival in lung cancer, identifies metagenes with unique spatial patterns linked to cell proliferation and immune response in kidney cancer, and detects genes preferentially expressed near amyloid-β plaques in an Alzheimer's model.

Spatial transcriptomics comprises a suite of genomic technologies enabling the profiling of gene expression on tissues with spatial localization information. Various spatial transcriptomics technologies have been developed, each with its distinct spatial resolution. Some technologies such as MERFISH[1–3] and Stereo-Seq[4] produce expression measurements on single molecules or with subcellular resolution, which, when paired with cell segmentation, can be further aggregated into single cells to achieve single-cell resolution. Other technologies such as 10x Visium[5] and Slide-Seq[6,7] operate at spot-resolution, collecting gene expression measurements at spots or beads corresponding to a mixture of cells with potentially heterogeneous cell types. Regardless of the resolution, various spatial transcriptomics technologies have been applied to diverse tissue types, revealing novel biological insights.

An essential analytic task in spatial transcriptomics involves identifying genes that display specific spatial expression patterns on tissues. These genes are known as spatially expressed genes (SE genes) or spatially variable genes (SVGs)[8–10]. SVGs represent transcriptomic signatures that underlie the topographical organization of complex tissues, shaping the spatial arrangement of tissue functions. However, many identified SVGs are cell-type marker genes, displaying spatial expression patterns that mirror the distribution of distinct cell types[11–15]. These SVGs can be readily identified through standard omics technologies such as single-cell transcriptomics. Consequently, the existing SVG analysis may offer limited value and potentially fail to fully capitalize on the potential of spatial transcriptomics.

Here, we advocate a paradigm shift in SVG analysis. Rather than detecting SVGs that are cell-type marker genes, we focus on identifying a subset of SVGs that display diverse spatial expression patterns within a specific cell type[10,16]. We refer to these SVGs as cell type-specific SVGs (ct-SVGs). ct-SVGs are different from differentially expressed (DE) genes or SVGs. Specifically, DE genes are genes that exhibit statistically significant differences in expression between pre-defined groups, where the groups may be cell types in single-cell data or spatial domains in spatial transcriptomics data[17]. SVGs, as explained above, are genes that display spatial expression patterns or are spatially variable across the entire tissue[8,10], which, by definition, includes DE genes that are differentially expressed between spatial domains.

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [3]Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [4]These authors contributed equally: Lulu Shang, Peijun Wu. ✉e-mail: xzhousph@umich.edu

ct-SVGs, on the other hand, represent another type of SVGs that display diverse spatial expression patterns or are spatially variable within a specific cell type[10,18]. Example ct-SVGs include *Aldoc*, which is highly enriched in a subset of Purkinje cells that are located in the posterior lobule of the cerebellum; as well as *Rasgrf1*, which is highly enriched in a subset of granule cells that are located in the anterior lobule of the cerebellum[19]. ct-SVGs can only be detected through spatial transcriptomics and have the potential to delineate the spatial transcriptomic heterogeneity within a particular cell type and are therefore crucial for understanding the transcriptomic mechanisms underlying cellular heterogeneity−how cells in the same cell type may exhibit different functional states, be influenced by local microenvironments, subject to distinct signaling molecules through communications[20,21], and go through the dynamic transcriptomic regulatory process underly cell state transitions during development[22] and disease progression[23,24]. Indeed, limited example ct-SVGs detected through ad hoc procedures have already begun to unveil complex developmental processes[25,26], serve as important biomarkers for clinical disease diagnostics[26,27], and yield invaluable insights into the differentiation of cell types, development, and organization of complex tissues, as well as disease progression and diagnosis[28]. However, despite its importance, there are currently no effective statistical solutions for detecting ct-SVGs (details in Supplementary Text).

To systematically identify ct-SVGs across a variety of spatial transcriptomics platforms, we present Celina (CELl type-specific spatially variable gene IdentificatioN Analysis), a statistical method that uses a spatially varying coefficient model to explicitly and accurately model gene's spatial expression pattern in relation to the cell type distribution across tissue locations. Celina provides effective type I error control and higher statistical power in detecting ct-SVGs in both single-cell and spot-resolution spatial transcriptomics. We applied Celina to five datasets, where it identified a range of ct-SVGs critical to tumor progression and invasiveness in human lung cancer, uncovered metagenes with distinct spatial patterns and functions in regulating the heterogeneous tumor cell proliferation and immune responses in human kidney cancer, and detected genes with preferential expression near amyloid-β plaques in hippocampus of an Alzheimer mouse model. The ct-SVGs detected by Celina open doors for novel biologically informed downstream analyses, unveiling functional cellular heterogeneity at an unprecedented scale.

Results
Overview and simulations
Celina is described in Methods, with its technical details provided in Supplementary Text and its method schematic shown in Fig. 1. We first performed extensive simulations to examine the effectiveness of Celina and compared it with several approaches that we adopt in the present study for ct-SVG detection. Simulation and method comparison details are provided in "Methods". In total, we explored 8 null simulation settings to examine type I error control and 40 alternative simulation settings across three distinct cell type-specific expression patterns (gradient, streak, and hotspot) to examine power for different methods. Because some methods failed to control for type I error, we measured power in the alternative simulations based on a fixed value of false discovery rate (FDR) to ensure fair comparison among methods.

Besides Celina, we adapted two sets of methods originated from other analytic settings to detect ct-SVGs. The detailed adaptations are outlined in Methods, with rationale described in Supplementary Text. The two sets of methods include those that were originally designed for detecting SVGs but are modified here to detect ct-SVGs (SPARK, SPARK-X), and those that were originally designed for other analytic tasks but are adapted here to detect ct-SVGs (CSIDE). For single-cell resolution simulations, we applied SPARK and SPARK-X to detect ct-SVGs in two different ways: we either applied them directly on the cells

extracted from the target cell type (SPARK_extract, SPARK-X_extract) or applied them on all cells and included the cell type labels as covariates (SPARK-X_control). We also applied CSIDE and Celina to detect ct-SVGs in single-cell resolution data using the cell type labels as covariates (CSIDE, Celina). For spot-resolution simulations, we applied the SVG detection methods directly on all spots (SPARK, SPARK-X) or applied them while adjusting for the true cell type proportions as covariates (SPARK-X_control). We also applied CSIDE and Celina using either true cell type proportions (CSIDE_oracle, Celina_oracle) or estimated cell type proportions using RCTD (CSIDE_wRCTD, Celina_wRCTD) or CARD (CSIDE_wCARD, Celina_wCARD) where we directly used the simulated single cells to serve as the reference. We also included CTSV for comparison, which is designed for detecting ct-SVGs along a pre-defined spatial direction. Note that method modification deemed impractical, such as SPARK_control, is not included in the analysis (detailed reasons provided in "Methods"). It is important to reiterate that most of the existing methods were not initially designed for detecting ct-SVGs. Consequently, their limitations in certain settings should not be considered as a reflection of their effectiveness for their originally intended tasks.

We first examined the simulation results for single-cell resolution spatial transcriptomics data. Under the null simulations, Celina produces well-calibrated *p*-values across different simulation scenarios (Fig. S1). SPARK-X_extract and SPARK-X_control also perform well and achieve calibrated type I error control as one would expect. SPARK_extract has calibrated *p*-values in scenarios I and II where one cell type dominates each spatial domain but has slightly conservative *p*-values in scenarios III and IV where there are multiple dominant cell types in each spatial domain. In contrast, a large fraction of the *p*-values from CSIDE (56% across scenarios) are exactly one (Fig. S1), suggesting failures of CSIDE in analyzing these datasets. The failure of CSIDE in producing *p*-values other than one in a large fraction of genes is expected due to its complicated testing apparatus (details in Supplementary Text) and is also observed in the permuted null dataset in ref. 18, where 26%–40% of *p*-values were exactly one across replications (Fig. S2). CTSV also produces calibrated *p*-values across different scenarios in the main simulation settings (Fig. S3). Consistent results are observed regardless of the number of spatial locations or the level of expression dispersion (Figs. S4–S8) and are observed in the complete null setting where all genes are non-SVGs (Fig. S9).

In the alternative simulations, Celina is more powerful than the other methods in detecting the gradient and streak patterns and as powerful as them in detecting the hotspot pattern (Figs. S1, S10). Specifically, at an FDR cutoff of 0.05, Celina achieves an average power of 96%, 61%, and 53% across the four scenarios in detecting the gradient, streak and hotspot patterns, respectively (Fig. S1). The power of SPARK_extract is 77%, 40%, and 63%, respectively; while the power of SPARK-X_extract is 73%, 26%, and 36%. The higher power of SPARK_extract over SPARK-X_extract is likely due to its parametric modeling assumption. Additionally, the power of Celina is slightly lower compared to SPARK_extract in detecting the hotspot pattern, presumably because the count model paired with the Gaussian kernels in SPARK_extract could be more effective in capturing the circular shape spatial pattern compared to Celina which models normalized data. In contrast, SPARK-X_control does not perform well with average power being 5%, 3% and 16% across four scenarios. Celina also outperforms two variants of CTSV, including CTSV_x (power = 0.2%, 0.2%, and 36%) and CTSV_y (power = 82%, 57%, and 30% (Fig. S11). CSIDE failed in all settings of scenario I with failure rate being 100%. CSIDE also yielded almost zero power across scenarios II-IV in detecting the gradient (average power=0.07%) and streak (0.2%) patterns and low power in detecting the hotspot pattern (8%). Using 5 or 10 basis functions in CSIDE improves its performance but did not change its ranking (Fig. S11) and the power of CSIDE decreases when the number of basis functions increases. Consistent results are observed regardless of the
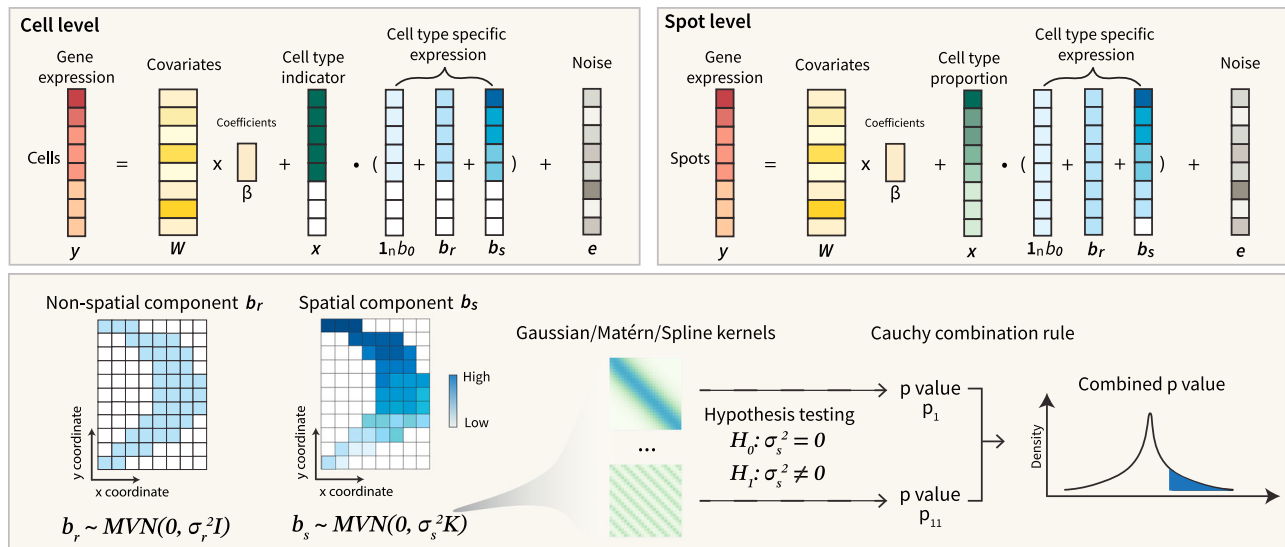
Fig. 1 | Overview of Celina. Celina is a method for identifying cell type-specific spatially variable genes (ct-SVGs) for both single-cell resolution (top left) and spot-resolution (top right) spatial transcriptomics. For a cell type of interest, Celina examines one gene at a time and takes as input the gene expression vector y, the cell type proportion (for spot-resolution data), or cell type indicator (for single-cell resolution data) vector x for the cell type of interest, any necessary covariates (W), and location information which is used for calculating the kernel matrix K. It relies on a spatially varying coefficient model to relate the gene's spatial expression pattern to the cell type proportions across tissue locations and uses random effects to disentangle the cell type specific expression b. We assume that b, apart from encompassing the mean, can be further decomposed into two additional components: the spatial component b_s, representing the part of cell type-specific gene expression explained by spatial correlation across locations; and the non-spatial component b_r, representing the part of cell type-specific gene expression not explained by spatial correlation across locations. Importantly, Celina relies on different kernel matrices K to capture a wide variety of cell type-specific expression patterns and outputs a combined p-value for each gene indicating the significance of the cell type-specific expression pattern.

number of spatial locations, the level of expression dispersion, or the strength of the spatial expression patterns (Figs. S12–S26).

Next, we examined the simulation results for spot-resolution spatial transcriptomics data. Under the null simulations, Celina_oracle retains its calibrated type I error control across scenarios (Figs. S27–S31). Celina_wCARD and Celina_wRCTD also maintain calibrated type I error control with estimated cell type compositions across different settings (Figs. S27–S31). As one might expect, SPARK and SPARK-X could not effectively distinguish SVGs from ct-SVGs for spot-resolution data and consequently produced p-values with inflated type I errors (Figs. S27–S31). Adjusting for the true cell type composition in the form of SPARK-X_control alleviated the p-value inflation issue to some extent, although inflation persisted in scenarios III and IV (Fig. S27). For CSIDE_oracle, an average of 47% of its p-values were exactly one under the baseline simulations across scenarios (Fig. S27). The non-one p-values from CSIDE_oracle remain inflated, with the genomic inflation factor ranging from 1.67 to 1.84 across scenarios. CSIDE_wCARD and CSIDE_wRCTD also result in a large faction of p-values equaling one (34% for CSIDE_wCARD and 49% for CSIDE_wRCTD). Their non-one p-values remain inflated, with genomic inflation factors ranging from 2.75 to 4.12 and 1.23 to 1.90 across scenarios, respectively (Figs. S27–S31). In addition, CTSV_x produced inflated p-values when it is paired with true cell type compositions, presumably because it falsely identified the cell type marker genes between spatial domains as ct-SVGs while CTSV_y produced calibrated p-values (Fig. S32). Similar results are observed across different levels of expression dispersion (Figs. S27–S31) and are observed in the complete null setting where all genes are non-SVGs (Fig. S33).

In the alternative simulations, Celina consistently outperforms the other methods across almost all settings (66 out of 68 across four scenarios) (Figs. 2d, S34–S48). Specifically, at an FDR cutoff of 0.05, when using the true cell type compositions, Celina_oracle achieves an average power of 99%, 87%, and 76% across the four scenarios in detecting the gradient, streak and hotspot patterns, respectively; while

the power of CSIDE_oracle is 0.2%, 0.3%, and 28% (Fig. 2d). Using 5 or 10 basis functions in CSIDE improves its performance but did not change its ranking (Fig. S49). Using relatively accurate cell type composition estimates from CARD only reduces the power of Celina and CSIDE slightly: Celina_wCARD achieves an average power of 98%, 76% and 56% across scenarios in detecting the three patterns; while the power of CSIDE_wCARD is 0.09%, 0.2%, and 4% (Fig. 2d). However, using the relatively inaccurate cell type composition estimates from RCTD can reduce the power of Celina further: Celina_wRCTD achieves an average power of 86%, 57% and 59% across scenarios in detecting the gradient, streak and hotspot patterns, respectively; while the power of CSIDE_wRCTD is 0.4%, 0.2%, and 18% (Fig. 2d). SPARK and SPARK-X achieves almost zero power (average power = 0%, 0%, and 0.05% in detecting the three patterns for SPARK, and = 0.4%, 0.7%, and 0.5% for SPARK-X), as they are unable to differentiate ct-SVGs from SVGs with the top genes containing a substantial number of false signals (Fig. 2d). On the other hand, SPARK-X_control achieves an average power of 26%, 8%, and 40% across the four scenarios in detecting the three patterns (Fig. 2d). Celina also outperforms two variants of CTSV when pairing with true cell type compositions, including CTSV_x (power = 0%, 0.1%, and 8%) and CTSV_y (power = 20%, 19%, and 10%) (Fig. S49). Consistent results are observed regardless of the level of expression dispersion, or the strength of the spatial expression patterns (Figs. S34–S48).

Finally, Celina performs scaling on the spatial coordinates during the kernel construction step, effectively converting the rectangular shape of the tissue section into a square (Methods). This scaling improves type I error control, particularly in spot-resolution simulations (Fig. S50), and enhances power (Figs. S51, S52), presumably because scaling improves algorithm stability and brings locations closer together to facilitate the detection of spatial expression patterns. Additionally, while Celina uses Scater[29] by default for data normalization, using an alternative normalization method SCTransform[30] achieves similar results (Fig. S53).
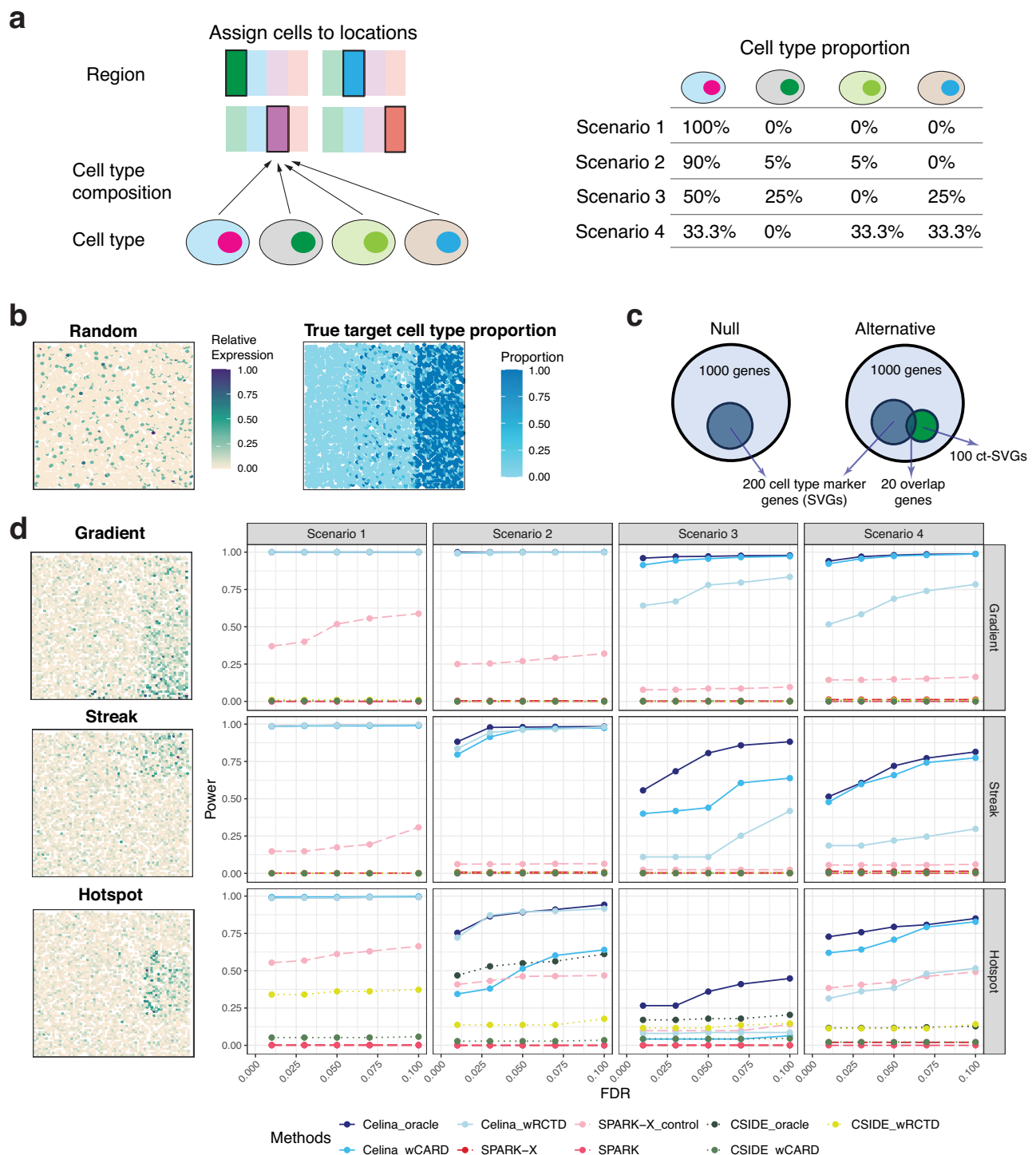
Fig. 2 | Simulation results for spot-resolution spatial transcriptomics data. **a** Simulation scheme. The simulated tissue contains four distinct rectangular spatial domains, each consisting of four distinct cell types. The composition of cell types in each spatial domain is varied to create four different simulation scenarios. **b** Left: a representative null gene displays random spatial expression pattern in simulation scenario II. Right: true cell type proportions are displayed on the tissue for the cell type of interest in simulation scenario II. **c** Left: each null simulation contains 1000 genes, among which 200 are cell type marker genes/SVGs (200). Right: each alternative simulation contains 1000 genes, among which 180 are cell type maker genes/SVGs but not ct-SVGs while 100 are ct-SVGs. **d** Left: representative genes display the three spatial expression patterns (Gradient, Streak and Hotspot) in simulation scenario II. Right: power plots show the proportion of true positives (y-axis) detected by different methods at a range of FDR (x-axis) in the alternative simulations for different methods (represented by different colors and line types) across the three spatial expression patterns. Simulations were performed under baseline setting. Source data are provided as a Source Data file.

Celina enables powerful ct-SVGs analysis in real data applications

Human lung cancer 10x Visium data. We applied Celina to analyze five published datasets obtained using different spatial transcriptomic technologies that include 10X Visium, STARmap-PLUS and Stereo-Seq (details in Methods). We only applied SPARK and SPARK-X to single-cell resolution datasets (last section in Results), omitting spot-resolution datasets due to the methods' inability to distinguish
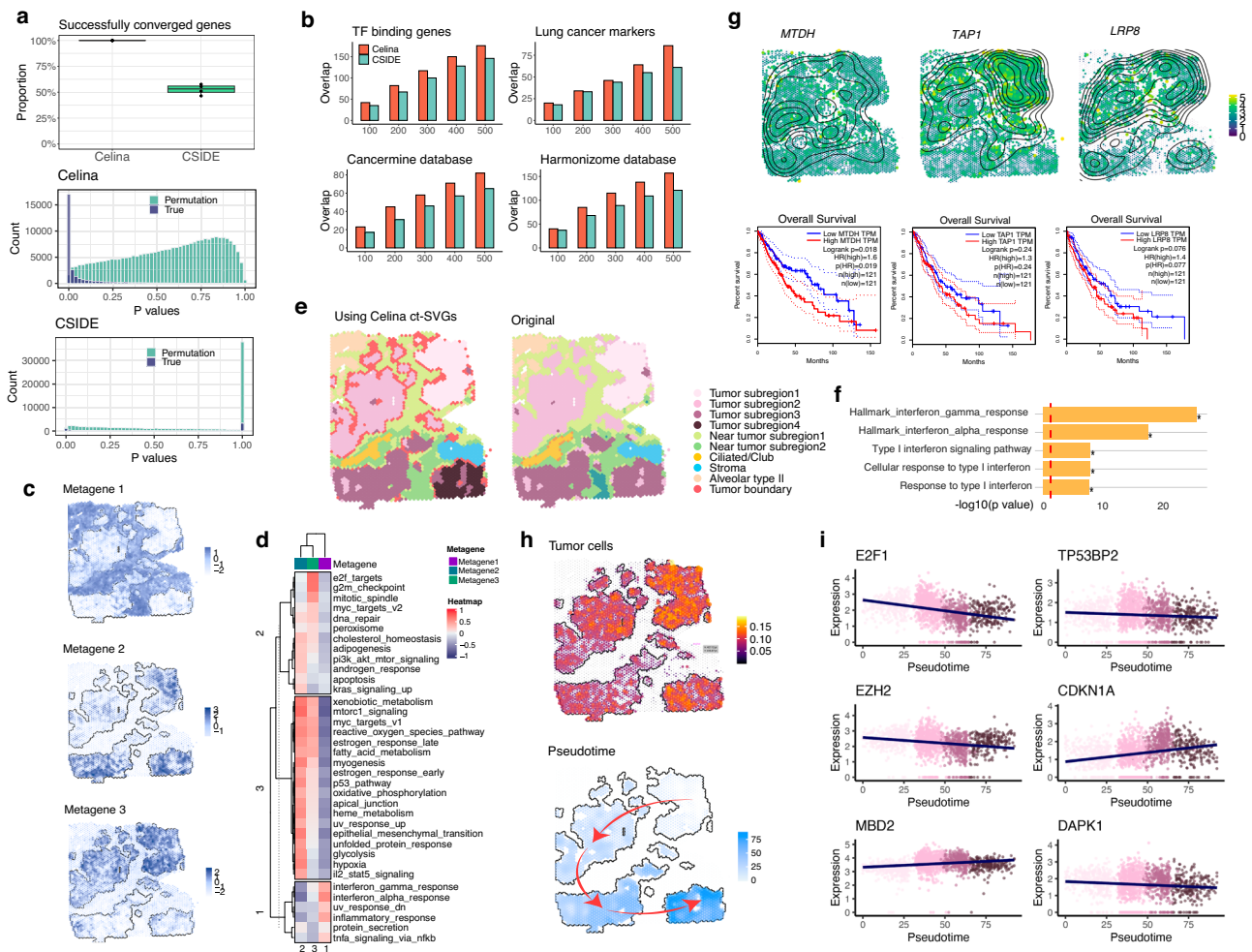
Fig. 3 | Results on the human lung cancer data. a Comparison of algorithm convergence rates and *p*-value distributions for Celina and CSIDE (*n* = 3813 spots). Top: boxplot shows the proportion of genes that achieved convergence in Celina and CSIDE. Middle: histogram displays the observed *p*-values from Celina in the permuted (green) and real data (purple). Bottom: histogram displays the observed *p*-values from CSIDE in the permutated (green) and real data (purple). In the box-plot, the center line, box limits and whiskers denote the median, upper, and lower quartiles, and 1.5 × interquartile range, respectively. **b** The number of overlaps (y-axis) between the top genes identified by Celina and CSIDE and known functional genes in existing databases (four panels), for top 100, 200, 300, 400, and 500 genes (x-axis). **c** Analysis of metagenes in tumor ct-SVGs. Tumor ct-SVGs identified by Celina are classified into three groups based on their distinct spatial expression patterns. The mean gene expression of each group is represented by a metagene displayed on the tissue. **d** Heatmap shows Pearson's correlation between hallmark pathway scores (rows) and metagene expression (columns). **e** Detection of tumor

boundary using ct-SVGs by Celina. Left: Spatial domains detected using ct-SVGs by Celina, with an important tumor boundary region displayed by red color. Right: Spatial domains detected using SVGs missed the tumor boundary. **f** The top 5 gene sets identified in the gene set enrichment analysis on genes differentially expressed in the tumor boundary region. Fisher's one-tailed test is used for functional enrichment analyses and the default option g:SCS method in gProfiler2 is used for multiple testing correction. **g** Spatial expression pattern for three tumor ct-SVGs, *MTDH*, *TAP1* and *LRP8*, identified by Celina (upper). Survival analysis using *MTDH*, *TAP1* or *LRP8* in TCGA, with *p*-values calculated using a two-sided log-rank test (bottom). **h** Trajectory analysis on tumor region. Top: visualization of tumor cell type proportion. Bottom: tumor stage trajectory in tumor region. **i** Expression level (y-axis) of transcription factors (Left) and their target genes (Right) is visualized along the tumor stage pseudotime (x-axis). Color represents four tumor subregions and solid line is fitted through linear regression. Source data are provided as a Source Data file.

ct-SVGs from SVGs, which leads to heavily inflated type I error as observed in simulations. We did not apply CTSV in the real datasets because of its heavy computational burden.

The first dataset was a human lung cancer data focusing on squamous cell carcinoma, which included 18,066 genes measured across 3858 spots. We performed ct-SVGs analysis in each of the 14 cell types that include natural killer (NK) cell, macrophage, endothelial cell, T cell CD4, B cell, tumor cell, T cell regulatory, stromal cell, plasma cell, T cell CD8, macrophage alveolar, transitional club/alveolar type 2 (AT2), alveolar type 1, and ciliated cell (Methods). We used a single cell reference data on lung cancer[31] to serve as the reference for deconvolution.

In the analysis, Celina produced slightly conservative *p*-values under the permuted null, with the genomic inflation factor ranging

from 0.48 to 0.65 for the 14 cell types (Fig. S54A). In contrast, CSIDE was only able to produce *p*-values under the permuted null for four cell types (plasma cell, stromal cell, transitional club/AT2, and tumor cell) (Fig. S54B). Even in these four cell types, CSIDE failed to converge in 0.03%–3.8% of the genes and erroneously produced a *p*-value exactly equaling to one in 34%–48% of the remaining ones (Fig. 3a), which is consistent with simulations. For the 46%–58% genes that CSIDE successfully run, its *p*-values remain inflated under the permuted null, with the genomic inflation factor ranging from 1.63 to 1.92 across the four cell types. Similar results were observed when 5 or 10 basis functions were used in CSIDE, which were only able to produce *p*-values for three cell types. At an empirical FDR of 0.01, Celina identified a total of 4423 ct-SVGs across 14 cell types, with 165–1989 genes per cell type, and achieved similar results when SCTransform was used for data

normalization (Fig. S55). In comparison, CSIDE detected 453 significant genes across four cell types, with 2-438 detected genes per cell type (Fig. S56), with more ct-SVGs than CSIDE across a range of FDRs, including its 5 basis function and 10 basis function variants (Fig. S57), even when restricting the comparison to genes and cell types that CSIDE was able to analyze (Fig. S58).

We evaluated the ct-SVGs detected by Celina, CSIDE by overlapping them with known functional genes in lung cancer. First, we obtained a list of 343 cell type marker genes from the original study[31] ("Methods") and found that a higher proportion of genes in the list are detected by Celina (307 out of 343) as compared to CSIDE (55 out of 343; Fisher's exact p-value < 2.2e-16). Second, we obtained a list of 932 lung cancer-related genes in Harmonizome database[32] and found that a higher proportion of genes in the list are detected by Celina (803 out of 932) as compared to CSIDE (121 out of 932; p-value < 2.2e-16). Third, we obtained a list of 542 cancer related genes including cancer drivers, oncogenes, and tumor suppressors from CancerMine database[33] and found that a higher proportion of genes in the list are detected by Celina (459 out of 542) as compared to CSIDE (65 out of 542; p-value < 2.2e-16). Finally, we obtained a list of 1113 transcription factors and their target genes and found that a higher proportion of genes in this gene list are detected by Celina (927 out of 1113) as compared to CSIDE (146 out of 1113; p-value < 2.2e-16). Celina also detected more functional genes than CSIDE among the top rank genes (Fig. 3b). The overlap results collectively support the higher power of Celina.

We carried out functional enrichment analysis on the ct-SVGs detected by different methods. We found that more pathways are enriched in genes detected by Celina (Fig. S60) than those detected by CSIDE (Fig. S61) and many of the enriched pathways for Celina are related to the biological functions in each cell type. Specifically, the top pathways in Celina include essential cancer related pathways such as MYC_Targets_v1 (p-value = 4.75e-42), E2F_Targets (p-value = 3.31e-33) for tumor cells; epithelial mesenchymal transition pathway (p-value = 6.7e-34) for plasma cells[34]; MTORC1 signaling pathway (p-value = 5.07e-27) for regulatory T cells, which is associated with homeostasis, activation and differentiation of T regulatory cells in tumor[35]; and critical metabolic pathways such as oxidative phosphorylation (p-value = 5.41e-16), respiratory electron transport (p-value = 1.82e-9) and ATP synthesis coupled electron transport (p-value = 6.11e-9) for Alveolar cell type 1[36]. In contrast, the top pathways in CSIDE are relatively generic and include hypoxia (p-value = 9.6e-7) for tumor cells; classical antibody-mediated complement activation (p-value = 0.62) for plasma cells; regulation of G protein-coupled receptor signaling pathway (p-value = 0.03) for stromal cells; and sequestering of actin monomers (p-value = 0.02) for transitional club/AT2 cells. This analysis further supports the superior performance of Celina in detecting genes enriched in biologically relevant pathways in the context of cancer.

We categorized the tumor ct-SVGs detected by Celina into three gene groups based on their spatial expression patterns and summarized the mean expression level in each group into a metagene. (Fig. 3c). The three gene groups display not only unique spatial expression patterns but also distinct biological functions (Figs. 3d, S61). Specifically, the genes in the first group display spatial variations within the tumor regions and are highly expressed in non-tumor regions with strong enrichment in immune response pathways such as interferon-alpha (IFN-α) and interferon-gamma (IFN-γ) response. Genes in the second group are preferentially expressed in the bottom third and fourth tumor subregions and are enriched with cancer metastasis pathways such as p53, hypoxia, and glycolysis. Genes in the third group are preferably expressed in the upper first and second tumor subregions and are enriched in cell cycle pathways such as G2M checkpoint and E2F targets. These findings indicate that the upper tumor regions likely represent early stages of the tumor, characterized by active cell proliferation. In contrast, the lower tumor regions are

indicative of late-stage tumor development, exhibiting increased invasiveness and undergoing progression towards metastasis. The categorization of tumor ct-SVGs thus highlights the potential of Celina in revealing spatial and functional cellular heterogeneity underling lung cancer.

The ct-SVGs detected by Celina also facilitate a biologically informed approach to delineate the tumor microenvironment. Specifically, we relied on the ct-SVGs identified by Celina to perform spatial domain detection and detected a thin layer of tumor boundary surrounding the tumor region (Fig. 3e). This thin layer of tumor boundary was not detected when using SVGs detected by SPARK, suggesting important information contained in ct-SVGs. Due to the lack of ground truth of spatial domain labels in the data, we extracted spot-level image features from the corresponding H&E image using deep learning models and used these features for validation. The spatial regions detected using Celina genes were better explained by the image features (pseudo-R^2 = 0.51) compared to those detected using CSIDE genes (pseudo-R^2 = 0.48), suggesting that Celina genes provide more spatial signals necessary for accurately defining tissue structures. Indeed, while both Celina and CSIDE-detected genes were effective in identifying the general tumor boundary, the tumor regions detected with Celina were more consistent with the morphology observed in H&E image while the tumor regions detected with CSIDE genes tended to display over-segmentation with the third tumor subregion divided further into two regions without clear H&E image support (Fig. S62). Notably, the genes differentially expressed in the identified tumor boundary are enriched in interferon-γ response pathway, which plays a key role in activation of cellular immunity and stimulation of antitumor immune-response[37] (Fig. 3f). The ct-SVGs detected by Celina also have a higher overlap with the differentially expressed genes as compared to CSIDE (Fig. S63). Example genes identified as ct-SVGs by Celina but missed by CSIDE (Fig. S64) include GBP1, a prognostic biomarker for lung cancer that encodes Guanylate Binding Protein 1. GBP1 is associated with immune cell infiltration and the development of tumor microenvironment in lung cancer[38] and is associated with survival outcomes in lung adenocarcinoma patients[39]. Therefore, the enrichment of the functional genes in the ct-SVGs detected by Celina also highlights its power in characterizing the biology underlying the detected spatial domains.

We carefully examined the ct-SVGs identified by Celina to reveal additional novel biological insights. For tumor cells, Celina detected 1247 ct-SVGs while CSIDE detected 438 (Fig. S56). Among them, Celina identified MTDH to be enriched in the second tumor subregion, TAP1 and LRP8 to be enriched in the first tumor subregion (Fig. 3g), all of which are missed by CSIDE. MTDH encodes the Metadherin protein, which is associated with dysregulated tumor immune environment and checkpoint expression[40]. TAP1 encodes a transporter protein, involved in the processing and presentation of MHC I restricted antigens including those associated with tumor[41]. LRP8 encodes a cell surface receptor, which facilitates the proliferation and invasion of non-small cell lung cancer cells by regulating the Wnt/β-catenin signaling pathway[42]. Importantly, we found that high expression levels of MTDH, TAP1 and LRP8 are associated with worse survival times in lung cancer patients using the external TCGA study (Fig. 3g). The detection of these genes suggests that they might serve as important markers for early diagnosis and prognosis of tumor.

Among the tumor ct-SVGs identified by Celina, notable ones include tumor-specific transcription factors such as E2F1, EZH2, and MBD2, along with their associated target genes. These genes are expressed along the inferred tumor stage trajectory, which points from early-stage tumor region that undergoes cell proliferation to the late-stage tumor region that undergoes metastasis and invasion (Fig. 3h, i). E2F1 is a member of the E2F family of transcription factors which play a crucial role in the control of cell cycle and activation of tumor suppressor proteins[43]. The enrichment of E2F1 in the early-stage

tumor region compared to the late-stage tumor region is consistent with its functional role in inducing expression of *TP53BP2*, which encodes tumor suppressor p53-binding protein 2 and is inversely correlated with the invasiveness of lung cancer[44]. On the other hand, *EZH2*, which encodes enhancer of zeste homolog 2, interacts with *UCA1* to epigenetically silence *CDKN1A* expression, contributing to drug resistance in lung cancer[45]. Consequently, *CDKN1A* expression increases as *EZH2* expression decreases from early-stage to late-stage tumor regions. As another example, *MBD2* encodes Methyl Cytosine Binding Domain Protein 2, which binds to the hyper-methylated promoters of tumor suppressor gene *DAPK1* to silence its expression and promote cancer progression. *MBD2* also displays gradual increase in expression along the tumor stage trajectory. These important tumor-specific transcription factors are only identified by Celina.

Besides tumor cells, Celina detected ct-SVGs in many other cell types (Fig. S65). For example, *NAMPT* is a transitional club/AT2 specific ct-SVGs detected only by Celina. *NAMPT* encodes Nicotinamide phosphoribosyltransferase, crucial for maintaining NAD+ levels in AT2 cells, and its upregulation by mesenchymal stromal cells helps reduce AT2 cell senescence, which is important for mitigating pulmonary fibrosis[46]. Increased *NAMPT* expression is associated with cancer metastasis[47,48]. As other examples, Celina is the only method that detected *C1QA* as a macrophage-specific ct-SVG. *C1QA* can promote an anti-inflammatory M2-like polarization in macrophages, thus contributing to an immunosuppressive environment[49]. In addition, Celina is the only method that detected *CDH5* as an endothelial ct-SVG. *CDH5* encodes VE-cadherin, a key adhesion molecule in endothelial cells, contributing to endothelial adherens junction assembly and maintenance[50,51]. The increased *CDH5* expression is associated with increased angiogenesis in lung cancer cells, promoting cancer cell migration and invasion[52]. These important genes are again only identified by Celina.

Human kidney cancer 10x Visium data. The second and third datasets we analyzed were collected on human kidney cancer. The two datasets contain two tissue sections from the same patient, one from tumor core with 36,601 genes measured across 3008 spots and one from tumor interface with 36,691 genes measured across 2048 spots. We performed ct-SVGs analysis in both tissue sections for each of the ten cell types that include B cell, plasma cell, endothelial cell (EC), renal cell carcinoma cell (RCC), non-proximal tubule epithelial cell (Epi_non-PT), proximal tubule epithelial cell (Epi-PT), fibroblast cell, myeloid cell, plasmacytoid dendritic cell (pDC), and mast cell (Methods). The cell type proportions used as input in Celina are obtained by using a single cell data collected on the same patient described in the original paper to serve as the reference data[53] for deconvolution. In the analysis, Celina produced slightly conservative *p*-values under the permuted null, with the genomic inflation factor ranging from 0.46 to 0.75 in tumor core and 0.3 to 0.66 in tumor interface across ten cell types (Fig. S66). In contrast, CSIDE was only able to produce p-values under the permuted null for two cell types (myeloid cell and RCC cell) in tumor core and three cell types (EC cell, plasma cell and RCC cell) in tumor interface. Even in these cell types, CSIDE failed to converge in 12–30% of the genes in tumor core and 16–26% in tumor interface (Fig. 4a). CSIDE erroneously produced a *p*-value exactly equaling to one in 19–22% of the remaining genes in tumor core and 25–35% of the remaining genes in tumor interface. For the 10–63% genes in tumor core and 54–54.2% genes in tumor interface that CSIDE successfully run, its *p*-values remain inflated under the permuted null, with the genomic inflation factor ranging from 3 to 3.9 across the two cell types in tumor core and 1.8–3.0 across three cell types in tumor interface.

At an empirical FDR of 0.01, Celina identified a total of 1769 ct-SVGs across ten cell types in the tumor core, with 58-745 genes per cell type (Fig. S67), and achieved similar results when SCTransform was

used for data normalization (Fig. S68). It also identified genes across ten cell types in the tumor interface, with 67-451 genes per cell type (Fig. S69). In comparison, CSIDE detected only 308 significant genes across two cell types in the tumor core, with 141–231 genes per cell type; and detected 111 significant genes across three cell types in the tumor interface, with 5–58 genes per cell type. Celina identified more ct-SVGs than CSIDE across a range of FDRs, even when restricting the comparison to genes and cell types that CSIDE was able to analyze. The results again support the superior power of Celina.

We evaluated the ct-SVGs detected by Celina and CSIDE by overlapping them with known functional genes in kidney cancer. First, we obtained a list of 371 kidney cancer-related genes in Harmonizome database[32] and found that a higher proportion of genes in the list are detected by Celina (247 out of 371 in tumor core; 216 in tumor interface) as compared to CSIDE (53 in tumor core, with Fisher's exact *p*-value < 2.2e-16; 33 in tumor interface, *p*-value < 2.2e-16). Second, we obtained a list of 69 kidney cancer related genes include cancer drivers, oncogenes, and tumor suppressors from CancerMine database[33] and found that a higher proportion of genes in the list are detected by Celina (41 out of 69 in tumor core; 35 in tumor interface) as compared to CSIDE (11 in tumor core, *p*-value = 9.8e-8; 5 in tumor interface, *p*-value = 7.5e-9). Finally, we obtained a list of 732 transcription factors and their target genes and found that a higher proportion of genes in this gene list are detected by Celina (460 out of 732 in tumor core; 387 in tumor interface) as compared to CSIDE (98 in tumor core, *p*-value < 2e-16; 46 in tumor interface, *p*-value < 2e-16). In addition, Celina detected more functional genes than CSIDE among the top rank genes (Fig. 4b). The overlap results collectively support the higher power of Celina.

We carried out functional enrichment analysis on the ct-SVGs detected by different methods. We found that more pathways are enriched in genes detected by Celina (Fig. S70) than those detected by CSIDE (Fig. S71) and many of the enriched pathways for Celina were related to the biological functions of each cell type. Specifically, the top pathways in Celina include hypoxia (*p*-value = 2.1e-13 in tumor core and *p*-value = 5.8e-14 in tumor interface) for RCC cells, which is closely associated with cancer cell progression; cellular response to IFN-γ (*p*-value = 1.55e-7 in tumor core) for myeloid cells and antigen processing pathway (*p*-value = 1.34e-4 in tumor interface) for plasma cell; and cell junction organization (*p*-value = 3.11e-5 in tumor interface) in EC cells. The top pathways in CSIDE are rather generic and include formation of the ternary complex (*p*-value = 0.02) for RCC cells in tumor core, and protein targeting to endoplasmic reticulum (*p*-value = 1.1e-6) for RCC cells in tumor interface; and complement activation (*p*-value = 0.014) for plasma cells in tumor interface. This analysis further supports the superior performance of Celina in detecting genes enriched in biologically relevant pathways in the context of cancer.

We carefully examined ct-SVGs detected in the renal cell carcinoma (RCC) cells. RCC cells distribute across the entire tissue section with enrichment either on the right side of the tumor core or on the upper region of the tumor interface (Fig. 4c). Importantly, genes identified by Celina from the two tumor samples are highly consistent with each other (Fig. 4d), more so than those detected by CSIDE. For example, Celina detected 745 significant RCC specific ct-SVGs in the tumor core and 451 genes in the tumor interface, 387 of which are commonly detected in both tissue sections. In contrast, CSIDE detected 231 significant RCC specific ct-SVGs in the tumor core and 58 genes in the tumor interface, only 35 of which are commonly detected in both tissue sections. Example genes detected by Celina include *NDRG1*, which is enriched in the left region of the tumor core and upper region of the tumor interface; *RPL13*, which is enriched in the middle region of tumor core and lower region of tumor interface; and *TMEM45A*, which is enriched in the right side of tumor core and middle region of tumor interface (Fig. 4c). *NDRG1* encodes a stress-responsive protein, is upregulated by cell differentiation signals in cancer cells, and suppresses
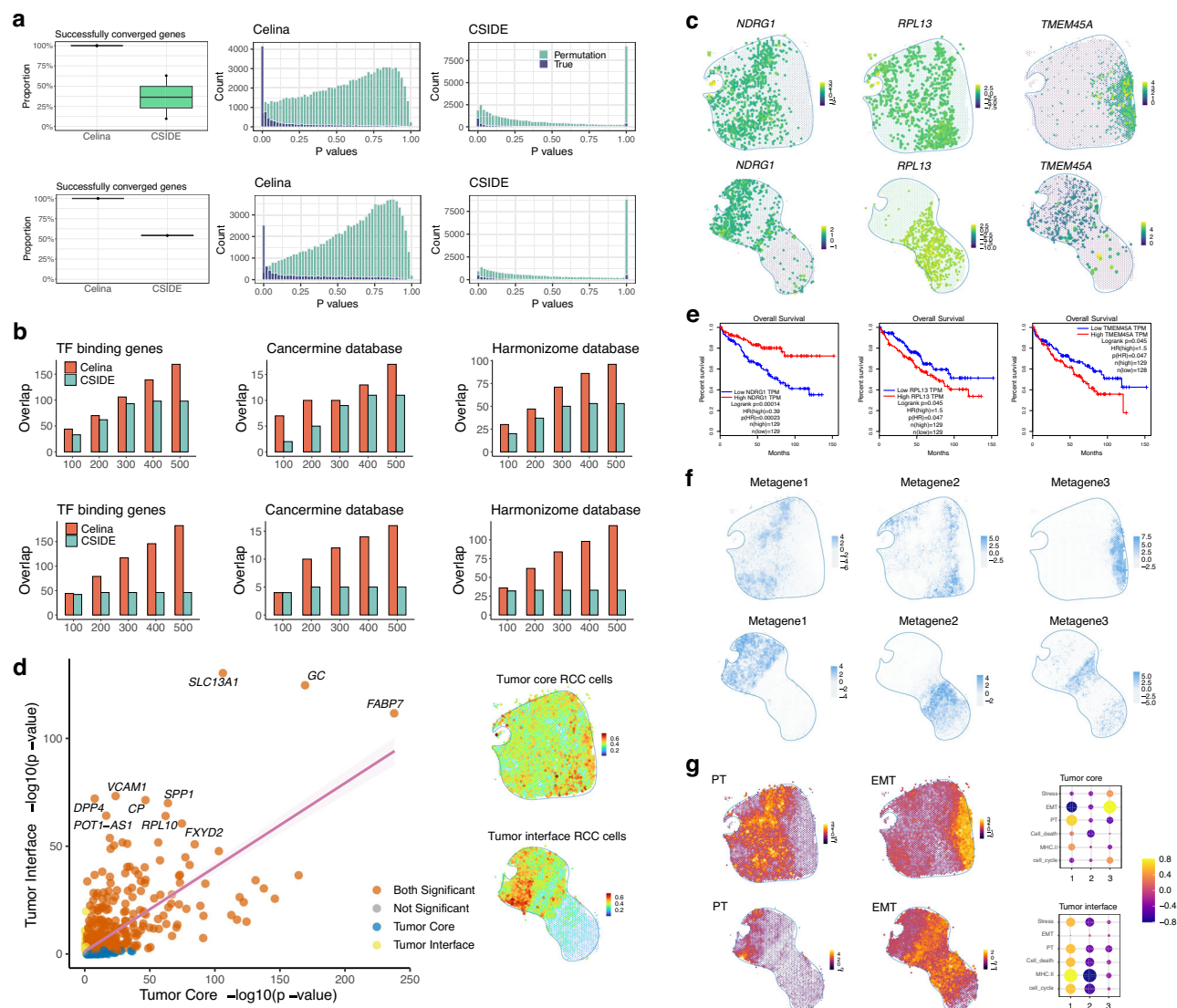
Fig. 4 | Results on the human kidney cancer data. a Comparison of algorithm convergence rates and *p*-value distributions for Celina and CSIDE (*n* = 2917 spots for tumor core and 2048 spots for tumor interface). Left: boxplot shows the proportion of genes that achieved convergence for different methods in the tumor core (upper) and tumor interface (lower). Middle: histogram displays observed *p*-values from Celina in permuted (green) and real data (purple). Right: histogram displays observed *p*-values from CSIDE in permutated (green) and real data (purple). In the boxplot, the center line, box limits and whiskers denote the median, upper, and lower quartiles, and 1.5 × interquartile range, respectively. **b** The number of overlaps (*y*-axis) between the top genes identified by Celina (red) or CSIDE (green) and known functional genes in public databases, for the top 100, 200, 300, 400, and 500 genes (*x*-axis). **c** Spatial expression pattern for three example ct-SVGs by Celina in RCC cells in the tumor core (upper) and tumor interface (lower). **d** Left: Scatter plot shows -log10(*p*-value) for RCC ct-SVGs in the

tumor core (x-axis) versus the tumor interface (y-axis). Each gene is labeled as red if it is significant in both datasets, grey if it is not significant in either dataset, blue if it is only significant in tumor core, and yellow if it is only significant in tumor interface. Right: Visualization of RCC cell type proportion in the tumor core (upper) and tumor interface (lower). **e** Survival analysis of the three ct-SVGs in RCC cells using TCGA, with *p*-value calculated using a two-sided log-rank test. **f** Analysis of metagenes in RCC cells in the tumor core (upper) and tumor interface (lower). Color represents relative metagene expression levels (blue for high; white for low). **g** Left: visualizing the proximal tubule (PT, left) and epithelial-to-mesenchymal transition (EMT, middle) meta-programs in the tumor core (upper) and tumor interface (lower). Right: Pearson's correlation between six meta-programs and metagenes in the tumor core (upper) and tumor interface (lower). Source data are provided as a Source Data file.

tumor metastasis[54]. *RPL13* encodes a ribosomal protein, whose high expression is associated with significantly poorer survival in kidney clear cell carcinomas[55]. *TMEM45A* encodes a transmembrane protein critical for tumor progression and resistance to chemotherapeutic agents in hypoxic condition[56]. Importantly, we found that the high expression level of *NDRG1* is associated with high survival rates in kidney cancer patients in the external TCGA study, while *RPL13* and *TMEM45A* are associated with low survival rates (Fig. 4e). These findings highlight the benefits of Celina in revealing the spatial and functional heterogeneity of kidney cancer.

We categorized the RCC-specific ct-SVGs detected by Celina into three gene groups based on their spatial expression patterns in tumor core and tumor interface, respectively, and summed their mean expression level in each group into a metagene (Fig. 4f). The three gene groups in each tissue section display not only unique spatial expression patterns but also distinct biological functions (Fig. S72). Genes in the first group are preferably expressed in the left region of the tumor core and the upper region of the tumor interface, both of which are enriched with interferon gamma (IFN-γ) response pathway that activates cellular immunity and stimulates antitumor responses in kidney

cancer[37]. The genes in the second group are preferably expressed in the middle region of the tumor core and the lower region of the tumor interface. These genes are enriched in the peptide chain elongation pathway and a large fraction of them (34% in the tumor core and 28% in the tumor interface) are ribosome genes. Ribosome is essential for protein synthesis, which is a fundamental process for cell growth and proliferation in cancer[57]. The genes in the third group are preferably expressed in the right region of the tumor core and the middle region of the tumor interface. These genes are enriched in the hypoxia and glycolysis pathways, both are critical hallmarks for tumor progression[58]. Notably, the first group of RCC ct-SVGs are also correlated with the proximal tubule program in the tumor core and MHC-II-related program in the tumor interface while the third group of genes are also correlated with epithelial-to-mesenchymal transition (EMT) in the tumor core (Fig. 4g). The categorization of ct-SVGs thus reveals novel biological insights into the cellular mechanisms underlying kidney cancer.

Mouse hippocampus STARmap-PLUS data. The fourth dataset was collected on a mouse hippocampus model of Alzheimer's disease (AD) at 13 months of age, containing 2766 genes measured across 10,372 cells ("Methods"). We performed ct-SVGs analysis in each of the 23 cell types that include interneuron, neuron, entorhinal cortex, deep layer subiculum, entorhinal cortex (IEGs), subiculum, dentate principal cells, medial entorhinal cortex, CA1 Principal cells, medial entorhinal cortexm, postsubiculum, CA2 Principal cells, CA3 Principal cells, astrocyte, oligodendrocyte, myelinProcesses, polydendrocyte, microglia, resident macrophage, ependymal, endothelial stalk, mural, and choroid_plexus (Methods). While this is a single-cell resolution spatial transcriptomics data, state-to-art cell segmentation methods are still limited in their ability to accurately define cell boundaries. Therefore, we first performed cell type deconvolution on the segmented cells by treating them as near-cellular spots, and then applied Celina and CSIDE to detect ct-SVGs for each of the 23 cell types. We used a single cell data collected in hippocampus from a public database for deconvolution (Methods). In the analysis, Celina produced slightly conservative p-values under the permuted null, with genomic inflation factor ranging from 0.7 to 0.99 across the 23 cell types (Fig. S73). In contrast, CSIDE was only able to produce p-values under the permuted null for one cell type (interneuron) (Fig. 5a). Even in this cell type, CSIDE failed to converge in 57% of the genes and erroneously produced a p-value exactly equaling to one in 29% of the remaining genes. For the 28% genes that CSIDE successfully run, their p-values remain inflated under the permuted null, with a genomic inflation factor of 2.96.

At an empirical FDR of 0.01, Celina identified a total of 71 ct-SVGs across 23 cell types, with 1–21 genes per cell type, and achieved similar results when SCTransform was used for data normalization (Fig. S74). In comparison, CSIDE detected only 1 gene in interneurons. Celina identified more ct-SVGs than CSIDE across a range of FDRs, even when restricting the comparison to genes and cell types that CSIDE was able to analyze (Fig. S75). We evaluated the ct-SVGs detected by Celina and CSIDE by overlapping them with 155 AD related genes based on GeneRIF Biological Term Annotations in the Harmonizome database[32]. Among these genes, Celina detected 5, while CSIDE detected 0 (Fig. 5b).

We carefully examined the ct-SVGs identified by Celina to reveal novel biological insights. We found that certain ct-SVGs exhibit a significantly higher expression level proximal to amyloid-β (Aβ) plaques (Fig. 5c). For example, *Vsnl1*, detected as a ct-SVGs in the entorhinal cortex (IEG), shows increased expression in entorhinal cortex cells located 10–20 µm from a plaque compared to those located >50 µm away (Wilcoxon-rank test p-value = 0.04; Fig. 5d). Similarly, *Lamp5*, identified in the postsubiculum, has higher expression in cells 0-10um from a plaque compared to those >50 µm away (p-value = 0.01). Additionally, *Tshz2* (p-value = 0.03), *Arf1* (p-value = 0.03), and *Nap1l5*

(p-value = 0.02), detected in the deep layer subiculum, all show higher expression in cells 10-20um from a plaque compared to cells >50 µm away. *Vsnl1* encodes visinin-like protein 1, influences synaptic plasticity and plays an important role in Aβ-induced and Ca2 + -mediated neuronal death in AD patients[59]. *Lamp5* encodes a lysosomal-associated membrane protein whose expression is reduced in AD brains and associated with neuronal network dysfunctions[60]. *Tshz2* encodes tea-shirt zinc finger homeobox 2, whose isoforms produce circTshz2-2, a circular RNA influencing spatial memory and neural structure and risks of neurological disorders including AD[61]. *Nap1l5* encodes nucleosome assembly protein 1-like 5 and is down-regulated in AD patients[62], with overexpression inhibiting expression of Aβ precursor protein in vitro[63]. These important ct-SVGs provide mechanistic insights towards AD and are only detected by Celina.

Besides the genes that preferably expressed near Aβ plaques, Celina detected ct-SVGs in other cell types (Fig. 5e). For example, *Timp2* is a marker gene for disease-associated-microglia (DAM), which are AD-associated phagocytic cells[64,65]. *Trf* encodes an iron-binding protein and is associated with late phase of oligodendrocyte maturation[66]. The C2 variant of *Trf* also plays a key role in mediating Fe-induced production of Aβ at the early stage of AD[67]. These important genes are again only detected by Celina.

Axolotl brain Stereo-Seq data. The fifth dataset was collected on axolotl brain, containing 22,893 genes measured across 4410 cells (Fig. 6a). We performed ct-SVG analysis in each of the six cell types that include CP (choroid plexus), EGC (ependymoglial cells), immature neuron, mature neuron, NBL (neuroblasts), and VLMC (vascular leptomeningeal cell; "Methods"). We were unable to perform cell type deconvolution in this data due to the lack of a single cell RNA-seq reference in axolotl brain. Consequently, we included SPARK_extract and SPARK-X_extract in this single-cell resolution data in addition to Celina and CSIDE to detect ct-SVGs for each of the six cell types (Fig. 6b). SVG analysis. In the analysis, Celina produced slightly conservative p-values under the permuted null, with the genomic inflation factor ranging from 0.84 to 0.93 across the six cell types (Fig. S76). The p-values from SPARK-X_extract under the permuted null were also conservative with a genomic inflation factor ranging from 0.78 to 0.89 across the six cell types, while the p-values from SPARK_extract were reasonably well calibrated with a genomic inflation factor ranging from 1.03 to 1.13. In contrast, CSIDE was only able to produce p-values under the permuted null for five cell types (except mature neuron) (Fig. S76). Even in these five cell types, CSIDE failed to converge in 0.03−0.54% of the genes (Fig. 6d) and erroneously produced a p-value exactly equaling to one in 23−44% of the remaining (Fig. 6b). For the 49−62% genes that CSIDE successfully run, their p-values remain inflated under the permuted null, with the genomic inflation factor ranging from 1.6 to 3 across the five cell types (Fig. 6c).

At an empirical FDR of 0.01, Celina identified a total of 2372 ct-SVGs across six cell types, with 42−1407 genes per cell type, and achieved similar results when SCTransform was used for data normalization (Fig. S77). SPARK_extract identified 2432 ct-SVGs across six cell types, with 78−1472 genes per cell type. SPARK-X_extract identified 2126 ct-SVGs across six cell types, with 140−1299 genes per cell type. In comparison, CSIDE detected only 1157 genes across five cell types, with 3−734 detected genes per cell type. Celina identified more ct-SVGs than CSIDE across a range of FDRs, even when restricting the comparison to genes and cell types that CSIDE was able to analyze (Figs. S78, S79). The ct-SVGs detected by the Celina, SPARK_extract and SPARK-X_extract are reasonably highly overlapped with each other, with Jaccard index ranging from 0.7 to 0.83 (Fig. S80). In contrast, the ct-SVGs detected by CSIDE have lower overlap with the other three methods, with Jaccard index ranging from 0.33 to 0.37 (Fig. S81). These results are consistent with the superior performance of Celina, SPARK_extract and SPARK-X_extract in single-cell resolution data as
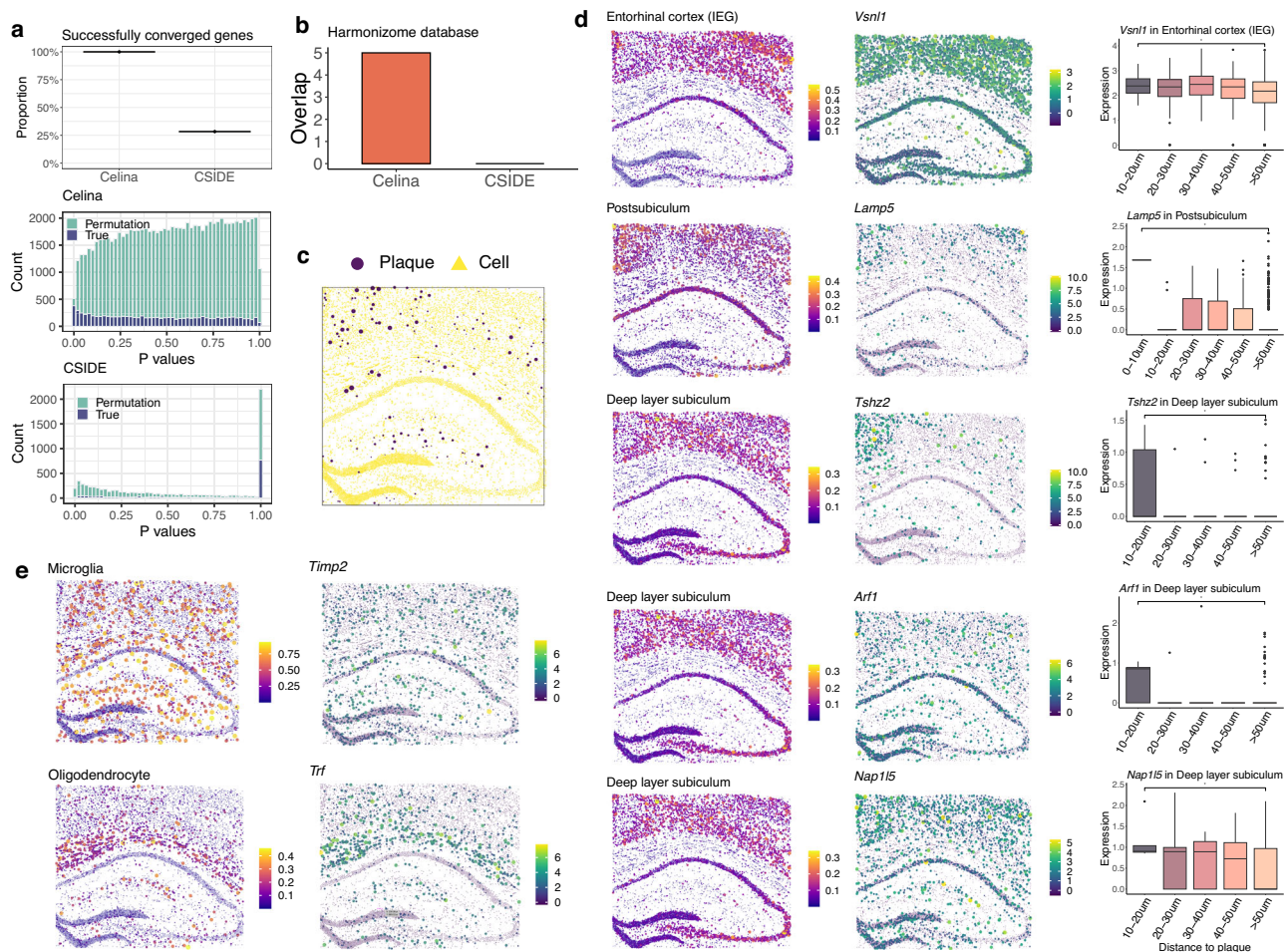
Fig. 5 | Results on the Alzheimer mouse hippocampus data. a Comparison of algorithm convergence rates and *p*-value distributions for Celina and CSIDE (*n* = 10,372 cells). Top: Boxplot shows the proportion of genes that successfully converged for Celina and CSIDE. Middle: Histogram displays the observed *p*-values from Celina in the permuted (green) and real data (purple). Bottom: Histogram displays the observed *p*-values from CSIDE in permuted (green) and real data (purple). In the boxplot, the center line, box limits and whiskers denote the median, upper, and lower quartiles, and 1.5 × interquartile range, respectively. **b** Number of overlaps between the top genes identified by Celina or CSIDE and known functional genes in public databases. **c** Visualization of Aβ plaques observed in the mouse Alzheimer's hippocampus data. Color represents plaques (purple) and cells (yellow). **d** Cell type proportion of five cell types (Left) and the spatial expression pattern of five ct-SVGs (Middle) on the tissue, with their expression stratified by distance to Aβ plaque (Right). Asterisks denote significantly different gene expression in the first distance interval compared to the >50 um interval. Two-sided Wilcoxon rank test is used with *p*-value being 0.04, 0.01, 0.03, 0.03 and 0.02 for five genes respectively, **p*-value < 0.05. In the boxplot, the center line, box limits and whiskers denote the median, upper, and lower quartiles, and 1.5 × interquartile range, respectively with the center line denoting the mean value of the expression. **e** Cell type proportion of two cell types (Left) and the expression level of two example ct-SVGs (Right) on the tissue. Source data are provided as a Source Data file.

observed in simulations. Certainly, SVG detection methods such as SPARK and SPARK-X detected a larger number of SVGs compared to ct-SVG specific methods focused on detecting a specific set of SVGs that are ct-SVGs (Fig. S82).

We evaluated the ct-SVGs detected by Celina and the other three methods by overlapping them with 419 cell type-specific marker genes[4]. We found that the top rank genes detected in Celina contain more marker genes than the other methods (Fig. 6d). We carefully examined the ct-SVGs detected by Celina and found that the detected genes revealed clear cell type-specific spatial expression heterogeneity. For example, Celina detected *Slc1a3* as a ct-SVG for ependy-moglial cells (EGC), which are equivalent to neural stem cells in mammals and reside in the ventricular zone for neurogenesis during development and regeneration[4]. *Slc1a3* is a neural stem cell marker in mouse[68] and axolotl[4] and is a top marker labeling two sub-populations of EGC (sfrpEGC and wntEGC)[4] (Fig. 6e, f). Consequently, the EGC specific spatial expression SVG pattern of *Slc1a3* identified by Celina likely reflects the spatial distribution of the two EGC sub-populations. Importantly, several marker genes in the EGC sub-populations, such as

Fabp7 in sfrpEGC and *Slit2* in wntEGC, exhibited similar spatial patterns compared to *Slc1a3*, providing additional support of evidence. Besides these genes, Celina also detected *Vax2* as a ct-SVG in dEGC, which is another sub-population of EGC. The spatial distribution of ct-SVGs within EGC subpopulations detected by Celina helps reveal the spatial and functional organization of EGC across brain regions.

Finally, we performed functional enrichment analysis on the ct-SVGs detected by each method. We found that the detected path-ways are more significant in the genes detected by Celina (Fig. S83), SPARK_extract (Fig. S84), and SPARK-X_extract (Fig. S85), than that detected by CSIDE (Fig. S86). The top GO terms in the genes detected by Celina include RNA binding in EGC (*p*-value = 1.07e-43), cyto-plasmic translation in NBL (*p*-value = 2.19e-23), cytosolic ribosome in CP (*p*-value = 1.48e-11) and VLMC (*p*-value = 7.07e-46), synapse in Immature_Neuron (*p*-value = 2.25e-25) and chemical synaptic trans-mission in Mature_Neuron (*p*-value = 0.1). Similar top GO terms are detected in SPARK_extract and SPARK-X_extract due to high ct-SVG overlap with that by Celina. In contrast, the top GO terms in genes detected in CSIDE included few significant pathways, including
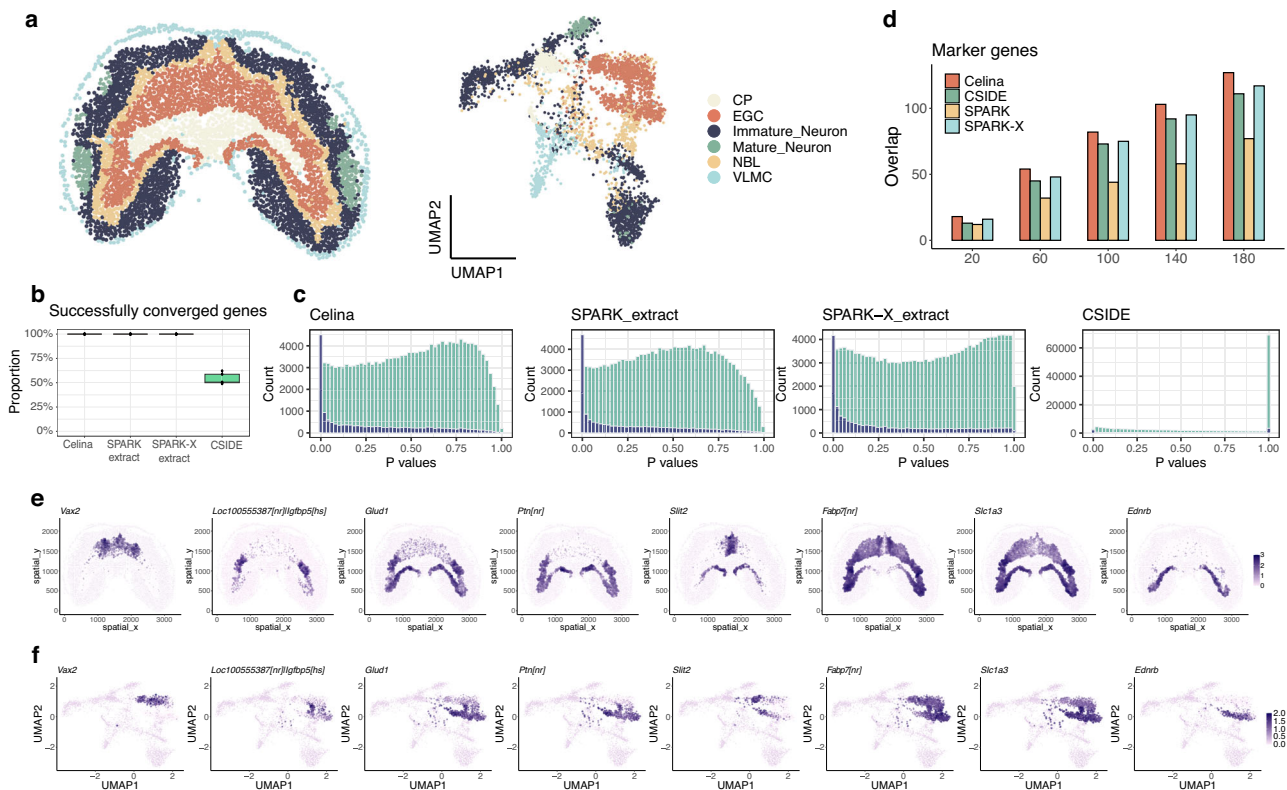
Fig. 6 | Results on the axolotl brain data. a Spatial distribution of cell types in the axolotl brain (Left) and on the UMAP (Right) for CP (choroid plexus), EGC (ependymoglial cells), Immature Neuron, Mature Neuron, NBL (neuroblasts), and VLMC (vascular leptomeningeal cell). **b** Boxplot shows the proportion of genes that successfully converged in different methods ($n = 4410$ cells). In the boxplot, the center line, box limits and whiskers denote the median, upper, and lower quartiles, and $1.5 \times$ interquartile range, respectively. **c** Histogram shows observed *p*-values from different methods in the permuted (green) and real data (purple). **d** Number of overlaps (y-axis) between the top genes identified by different methods and known marker genes, for the top 20, 60, 100, 140, and 180 top genes (x-axis). **e** Spatial expression pattern for example ct-SVGs identified by Celina in EGC cells. **f** UMAP plot for the same ct-SVGs. Source data are provided as a Source Data file.

cytosolic ribosome in EGC (*p*-value = 3.62e-29), neuron development in Immature_neuron (*p*-value = 6.72e-20) and nucleic acid binding in NBL (*p*-value = 3.74e-05). CSIDE did not identify significant GO terms in CP and VLMC. Many of the enriched gene sets identified by Celina were related to neuron development and differentiation in immature neuron cells and synaptic signaling in mature neurons. This analysis further supports the superior performance of Celina in detecting biologically relevant pathways in the context of brain development in axolotl.

Discussion

We have presented Celina, a statistical method for identifying ct-SVGs that display cell type-specific spatial expression patterns. Celina explicitly models the gene's spatial expression pattern in relation to the cell type distribution across tissue locations utilizing a spatially varying coefficient model, providing robust performance across spatial transcriptomics platforms with distinct resolutions. Celina is computationally efficient and scalable to data with tens of thousands of spatial locations and thousands of genes. Through detecting ct-SVGs, Celina facilitates our understanding of the spatial and functional cellular heterogeneity within tissues and across diseases.

In the real data, we found a sizable proportion of detected ct-SVGs are shared across cell types, with 28.6%-76.9% of these genes detected in at least two cell types across datasets (Fig. S87). While such overlap does not influence the comparison and conclusion in our study, it does highlight the complex spatial expression patterns these genes may exhibit in a cell type-specific manner. There are several potential biological explanations behind the observed overlaps. Some genes expressed across multiple cell types may regulate fundamental

biological processes or cellular functions essential for all those cell types, hence displaying cell type-specific expression patterns in multiple cell types. Some genes may contribute to the organization and formation of spatial domains on the tissue and thus display cell type-specific expression patterns in multiple cell types that coexist on these domains (e.g., sharp boundaries within layered tissue structures[69]). Some genes may display coordinated cell type-specific spatial expression patterns across multiple cell types due to intercellular interactions and influence from the surrounding microenvironment shaped by communication signals, nutrient gradients, oxygen availability, and more. Understanding these dynamics is essential for accurately understanding the cell type-specific spatial gene expression in spatial transcriptomics studies.

Celina applies scaling to the spatial coordinates before constructing the covariance matrix and performing subsequent analysis. Scaling spatial coordinates is a common practice in spatial transcriptomics data analysis, but it may distort the original spatial correlation—for example, creating a square-like shape if the original tissue section is rectangular. In the real datasets explored in the present study, the aspect ratio—defined as the ratio of width to height of the tissue section—is relatively mild, ranging from 0.57 to 1.51. Our simulation results indicate that the choice of scaling method has a negligible impact on tissue samples with aspect ratios between 0.5 and 2.0 (Figs. S50–S52). However, it is important to note that different scaling approaches could yield varying results in cases with more extreme aspect ratios. To facilitate practical applications, we have implemented multiple scaling approaches in Celina.

Celina is not without limitations. First, for spot-resolution spatial transcriptomics data, like ad hoc modifications for ct-SVG analysis,

Celina also requires cell type composition estimates across tissue locations obtained through cell type deconvolution. Consequently, the accuracy of Celina necessarily depends on the accuracy of cell type devolution. Indeed, as we have shown in the simulations, Celina achieves the highest power in detecting ct-SVGs when supplied with the true cell type compositions, slightly lower power with the estimated cell-type compositions from CARD, and yet lower with RCTD. Fortunately, the flexible modeling framework of Celina allows it to be paired with any devolution method as needed to take advantage of their benefits. Second, Celina uses normalized data as input, which ensures its computational efficiency. Modeling normalized data, however, may not be ideal as the raw spatial transcriptomics data are in the form of sparse counts with a prevalence of zero values and unique mean-variance relationship[70]. Consequently, extending Celina to direct model counts, in the generalized linear modeling framework, may have added benefits. However, direct modeling of large-scale count data may be computationally challenging, as a negative-binomial model or an over-dispersed Poisson model not only is slow but also can lead to a failure of convergence in more than 90% of genes in large-scale spatial transcriptomic data[10]. Therefore, generalizing Celina for count modeling while balancing computational efficiency and numerical stability is an important future direction. Third, Celina analyzes one gene and one cell type at a time. However, genes may interact with each other, are organized in biological pathways, and may display correlated expression patterns. Consequently, it would be beneficial to extend Celina to jointly model multiple genes together in order to borrow information across correlated genes to better characterize the spatial expression patterns in a cell type-specific manner. Fourth, our study is a computational methodology study that did not include wet-lab experimental validation, which is an important limitation commonly found in the majority of computational methodological studies. Finally, exploring the benefits of Celina to other spatial omics data, such as spatial epigenome data, could help characterize the spatial organization of chromatin accessibility across the tissue in a cell type-specific fashion.

Methods

Method overview

We aim to identify genes that display spatial expression patterns, known as spatially expressed genes (SE genes) or spatially variable genes (SVGs), in a specific cell type in spatial transcriptomics studies. We assume that a spatial transcriptomics study collected expression measurements on n different spatial locations on the tissue. These locations have known spatial coordinates recorded during the experiment. We denote s as an n by d matrix of spatial coordinates, where the ith row s_i is a d-dimension vector of spatial coordinates for location i, with $d = 2$ or 3 depending on the particular spatial transcriptomics technology and $i \in \{1, 2, ..., n\}$. In the spatial transcriptomics study, we examine one gene at a time and denote $y_i(s_i)$ as the normalized gene expression measurement for the given gene on the i-th location. We denote $y = (y_1(s_1), ..., y_n(s_n))^T$ as an n-vector of normalized expression measurements across all locations with mean centered to zero. We assume that there are k different cell types in the data and examine one cell type at a time. For the particular cell type of interest, we denote x as the n-vector of the cell type proportion across locations. In single-cell resolution spatial transcriptomics data, x is an n-dimension vector of binary cell type indicators, where the i-th element indicates whether the i-th cell belongs to the particular cell type or not. In spot or bead resolution spatial transcriptomics data, x is an n-dimension vector of cell type proportion, where i-th element represents the percentage of the cell type in the i-th location. The cell-type compositions for spot or bead resolution spatial transcriptomics data can be estimated through existing cell-type deconvolution methods. We primarily used CARD[71] to obtain the cell type proportion for the spot or bead resolution data in the present study.

With the above notations, we rely on a spatially varying coefficient model to relate the gene's spatial expression pattern to the cell type proportions across tissue locations. Specifically, we assume:

$$y = W\beta + x \cdot b + e, \tag{1}$$

where $\cdot$ represents a dot product; b is an n-vector of cell type-specific gene expression level across locations; xb is a n-vector of gene expression that is contributed by the cell type of interest; W is a $n \times (k-1)$ covariates matrix includes the cell type proportions for all other cell types; β is a $(k$-1)-vector of corresponding coefficients; intercept is not included in W since we have centered the response y as described above; $W\beta$ represent the gene expression that is contributed by all other cell types; and e is a n-vector of residual errors with each element following a normal distribution with mean zero and variance σ_e^2.

The cell type-specific gene expression across locations, b, is a key set of parameters in the model. Our goal is to examine whether b displays any spatial expression patterns and is thus spatially varying across locations. To do so, we first denote the scalar b_0 as the mean cell type-specific expression across locations. We assume that b, apart from encompassing the mean, can be further decomposed into two additional components in the form of $b = \mathbf{1}_n b_0 + b_s + b_r$: the n-vector of b_s represents the spatial component of b, which measures the part of the cell type-specific gene expression in b that is explained by spatial correlation across locations; while the n-vector of b_r represents the non-spatial component of b, which measures the part of the cell type-specific gene expression in b that is not explained by spatial correlation across locations. To ensure the identifiability of the decomposition of b, we assume a priori that

$$b_s = \left(b_{s_1}(s_1), ..., b_{s_n}(s_n) \right)^T \sim MVN(0, \sigma_s^2 K), \tag{2}$$

$$b_r = \left(b_{r_1}(s_1), ..., b_{r_n}(s_n) \right)^T \sim MVN(0, \sigma_r^2 I), \tag{3}$$

where MVN denotes a multivariate normal distribution; σ_s^2 and σ_r^2 are the variance components that effectively capture the contribution of spatial and non-spatial components to the cell type-specific expression, respectively; I is an $n \times n$ identity matrix that imposes a priori independence among the elements of b_r, thus ensuring the non-spatial nature of b_r; and K is an $n \times n$ kernel matrix that is constructed based on the distance matrix s to model the correlation in b_s across spatial locations, thus inducing spatial expression patterns in b_s. With the above modeling assumptions, $b(s) \sim MVN(\mathbf{1}_n b_0, \Sigma)$, where $\Sigma = \sigma_s^2 K + \sigma_r^2 I$. Therefore, b is a spatially varying coefficient term that builds upon a Gaussian process and varies across spatial locations. Because b_s captures the spatial component of the cell type-specific gene expression, detecting ct-SVGs is equivalent to testing whether the elements of b_s are all zero, which is equivalent to a variance component test of the null hypothesis $H_0 : \sigma_s^2 = 0$.

The statistical power of the above variance component test is expected to be dependent on the specification of the kernel matrix K. In particular, the power of the test is high if the spatial kernel K captures the cell type-specific spatial correlation pattern of the given gene but can be low otherwise. However, the true cell type-specific spatial correlation pattern of any gene in the real data is usually unknown and such spatial correlation pattern may also vary across different genes. To ensure the effectiveness and robustness of the variance component test, instead of using a single fixed kernel matrix, we take two alternative approaches depending on the data size. Specifically, for small datasets such as those obtained from 10x Genomics ($n \leq 8000$), we followed[10] and constructed a wide range of spatial kernel matrices (a total of 11) to capture a wide variety of possible cell type-specific

spatial correlation patterns that may be encountered in the real data. To do so, we first scaled the x coordinates across locations to have mean 0 and standard deviation of one and scaled the y coordinates across locations to also have mean 0 and standard deviation of one. We then obtained the Gaussian kernels in the form of $K_G(s_i, s_j) = \exp(-|s'_i - s'_j|^2/2\sigma^2)$, the Matérn kernels in the form of $K_M(s_i, s_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{2\nu}}{l}|s'_i - s'_j|)^\nu K_\nu(\frac{\sqrt{2\nu}}{l}|s'_i - s'_j|)$ and the spline kernel in the form of $K_S(s_i, s_j) = m'(m'^T m')^{-1}m'^T$. Here, σ denotes the bandwidth parameter in the Gaussian kernels; ν (= 1) denotes the smoothness parameter in the Matérn kernels; l denotes the scale parameter in the Matérn kernels; $\Gamma()$ denotes the Gamma function; K_ν denotes Bessel function; m denotes the spline functions (default is 4) obtained from *smoothCon* function in the "*mgcv*" R package[72–74]. The Gaussian kernels induce spatial correlation in the cell type-specific gene expression on the tissue, encouraging expression similarity in neighboring locations and therefore intend to capture the smooth spatial patterns. The bandwidth parameter in the Gaussian kernels determines the strength of spatial correlation: a large parameter leads to a high spatial correlation and vice versa. On the other hand, the Matérn kernels are more flexible and can therefore capture the rough spatial patterns. The scale parameter in the Matérn kernels determines the length scale of spatial correlation: a smaller value means that the function varies more quickly with distance, while a larger value means that it varies more slowly. In addition, the spline kernel is semi-parametric and can complement Gaussian and Matérn kernels in detecting complex, nonlinear spatial patterns. In the above Gaussian and Matérn kernels, we selected the bandwidth parameter σ and the scale parameter l by using the 20%, 40%, 60%, 80%, and 100% quantiles of the absolute values of the scaled x and y coordinates together in the analyzed data to capture a wide variety of possible spatial patterns[8]. However, we note that our implementation can easily handle many other kernel functions or incorporate a different number of kernel functions from the input. For large datasets ($n > 8000$), due to computational reasons, we use a single Gaussian kernel matrix in the form of $K(s_i, s_j) = \exp(-|s'_i - s'_j|^2/\gamma)$, where γ is the bandwidth parameter. Instead of fixing the bandwidth parameter γ, however, we determine γ in a data-specific fashion to ensure the effectiveness and robustness of the variance component test. Specifically, we first follow previous approaches[75–77] and obtain Silverman's "rule-of-thumb" bandwidth, which is defined as $0.9 \min(\hat{\sigma}, \frac{IQR}{1.34})n^{-\frac{1}{5}}$, for one gene at a time, where $\hat{\sigma}$ is the standard deviation of gene expression and IQR is the interquartile range of gene expression[76]. Afterwards, we obtain the median bandwidth value across all genes to serve as γ. After constructing the kernel matrix K for large dataset, we further apply a low-rank approximation on K, with rank set to be two, to make it consistent with the rank in the projection covariance function. This way, we ensure effectiveness of the kernel in capturing spatial expression patterns while keeping computation in check.

For each kernel constructed above, we fitted the null model, which is defined as Eqs. (1)-(3) but without the b_s term. During model fitting, we relied on the restricted maximum likelihood estimation[78] and utilized average information (AI) algorithm[79] to iteratively update the variance components and the other parameters in a computationally efficient fashion. Afterwards, we obtained the score test statistic, which is in a quadratic form of the gene expression vector y. Consequently, the null distribution of the score statistic follows a mixture of chi-square distributions[80], allowing us to apply the Satterthwaite method[81] on the score statistics to calculate the p-value. In the presence of the multiple kernels, we further use the Cauchy combination rule[82] to combine the p-values obtained from different kernels into a single, combined, p-value. The resulting p-value is used to serve as evidence for detecting genes with cell type-specific spatial expression patterns. In each real dataset, for each cell type, we permuted the spot/cell coordinates ten times to construct an empirical null distribution of p-values for each method. Note that each

permutation includes 2431–5696 genes in the real datasets, so performing ten permutations corresponds to 24,310–56,960 cases, which is sufficient to create a null distribution. Indeed, we explored a setting with 1000 permutation and found the results to be nearly identical to those with 10 permutations (Jaccard index for identified genes = 0.98). Based on the empirical null distribution of p-values, we then followed[8,10] and calculated empirical FDR for each p-value cutoff and declared a gene as significant based on an FDR threshold of 0.01.

We refer to our method as Celina (CELl type-specific spatIal patterN Analysis in spatial transcriptomics), which is implemented as an R package. We used efficient C/C++ codes linked through Rcpp and with multiple threads to achieve high computing capability. The software, together with all analysis code for reproducing the results in the present study, are freely available at https://xiangzhou.github.io/software/.

Simulation design

We performed comprehensive simulations to evaluate the performance of Celina and compared it to other methods for detecting ct-SVGs. Specifically, we followed[10,83] to simulate single-cell resolution and spot-resolution spatial transcriptomics data on the primary visual cortex tissue collected in ref. 84. For single-cell resolution spatial transcriptomics data, we first randomly generated the x and y coordinates for a fixed number of cells (n) through a random-point-pattern Poisson process following[85]. These cells belonged to four different cell types and were further assigned to four spatial domains that correspond to four cortical layers (L2-L6) following[85]. We assumed that each spatial domain contains multiple cell types, and we varied the composition of cell types in each spatial domain to create four different simulation scenarios. In scenario I, each spatial domain contains one unique cell type. In scenarios II-IV, each spatial domain consists of three cell types and the missing cell type in each domain is different from each other. In scenario II, the proportion of the dominant cell type is set to be 90% and the proportion of each non-dominant cell type is set to be 5%. In scenario III, the proportion of the dominant cell type is set to be 50% and the proportion of each non-dominant cell type is set to be 25%. In scenario IV, the proportion of each of the three cell types is set to be equal (1/3). In each scenario, we assigned each cell to a cell type based on a categorical distribution, where the probability parameters in the distribution were set to be equal to the cell type composition in the spatial domain where the cell resided. With the cell type assignment, we simulated gene expression count data for 1000 genes in the cell using a negative binomial distribution, with the distribution parameters inferred from the cortex data[84]. For inferring these parameters in the cortex data, we first used the *fitdistr* function in the "*MASS*" R package[86] and fitted a negative binomial distribution for every gene. We then obtained the average of the mean parameters (μ) inferred across genes (= 1.0) to serve as the mean parameter for simulations. We also obtained the 25th, 50th, and 75th quantiles of the estimated dispersion parameters (θ) across genes (= 0.7, 0.95, and 1.3, respectively) to serve as the dispersion parameters for simulations. With these inferred parameters, we simulated expression count data for 1000 genes in each cell in both null and alternative simulations, with details described below. We randomly selected 200 genes among 1000 genes to be cell type marker genes with 50 unique marker genes per cell type. These cell type marker genes serve as differentially expressed (DE) genes to differentiate between the four cell types. Here cell type marker genes are genes that exhibit statistically significant differences in expression between cell types. We selected an equal proportion of the DE genes to be upregulated or downregulated: we set the fold change to be 4 for upregulated DE genes and 0.25 for the downregulated ones for cell types one and four, and set the fold change to be 2 for upregulated DE genes and 0.5 for the downregulated ones for cell types two and three.

In the null simulations, these 200 cell-type marker genes are SVGs due to the spatial distribution of the cell types, but none of them displays spatial variation in a cell type-specific fashion and are thus considered as null as they are not ct-SVGs. Besides the marker genes, we set the remaining 800 non-marker genes to be non-SVGs with random expression patterns across spatial locations. Therefore, the expected number of DE genes, SVGs, and ct-SVGs in the null simulation are 200, 200 and 0 respectively. In each null simulation scenario, we varied the number of single cells (n) on the tissue to be 3000, 5000, or 10,000. We varied the dispersion parameter (θ), which characterizes the noise level in the data, to be 0.7, 0.95, or 1.3. We set a baseline simulation setting with $n = 5000$ to ensure computational efficiency and $\theta=0.95$ and varied one parameter at a time on top of the baseline setting to create a total of five simulation settings per scenario, with ten simulation replicates per setting.

In the alternative simulations, we first randomly selected 200 cell-type marker genes, which are also SVGs. Afterwards, we randomly selected 100 genes among the 1000 genes to be ct-SVGs. Among these 100 ct-SVGs genes, on average, 20 were selected from the cell type marker genes while 80 were selected from the non-cell type maker genes. Therefore, in expectation, 180 genes are cell type marker genes/ SVGs but not ct-SVGs, while 100 genes are ct-SVGs which are also a type of SVGs, resulting in 280 SVGs in total. Consequently, the expected number of cell type marker genes, SVGs, and ct-SVGs in the alternative simulation are 180, 280, and 100 respectively. For the 100 ct-SVGs, we selected the cell type with the largest number of cells as the target cell type and simulated their expression in the target cells to display three distinct cell type-specific spatial expression patterns, namely hotspot, streak, and gradient patterns, detailed below.

For the hotspot pattern, we first created a circle around the geometric center of the cells from the target cell type and marked the cells of the target cell type within the area. We determined the radius of the circle such that a fixed proportion of cells (pi) from the target cell type were within the circle, with pi set to be either 0.3, 0.4 or 0.5. We then partitioned the circle vertically into two equal halves and simulated gene expression levels for the cells from the target cell type in each half of the circle separately in order to create cell type-specific spatial expression pattern. Specifically, for cells in the right half of the circle, we simulated their expression levels from a negative binomial distribution with mean set to be either 2 or 2.5 times higher than the mean of the cells in the target cell type that reside outside of the circle (for 50 ct-SVGs) or 2/5, 1/2 times lower than the mean of the cells in the target cell type that reside outside of the circle (for the remaining 50 ct-SVGs) for the 100 ct-SVGs. These fold change (f) values were selected to display either weak or strong ct-SVGs patterns, respectively. For cells in the left part of the circle, we simulated their expression levels from a negative binomial distribution with mean being 2 times higher (for 50 ct-SVGs) or 1/2 times lower (for the remaining 50 ct-SVGs) than the mean for the cells in the right half. For each simulation scenario, we set a baseline simulation setting with $\theta=0.95$, $f = 2$, $pi = 0.4$ and $n = 5000$. We then varied each of the four parameters ($\theta, f, prop, n$) one at a time on top of the baseline setting to create a total of eight simulation settings per scenario, with five simulation replicates per setting.

For the streak pattern, we created a rectangular stripe marking the upper part of the tissue for cells from the target cell type. We determined the stripe size such that a fixed proportion of cells (pi) from the target cell type are included within the stripe, with pi set to be either 0.1, 0.2 or 0.3. For the 100 ct-SVGs, we simulated the expression levels for cells of the target cell type inside the stripe from a negative binomial distribution, with mean being either 1.5 or 2.0 times higher than the mean of the remaining cells of the target cell type (for 50 ct-SVGs) or 1/2 or 2/3 times lower than the mean of the remaining cells of the target cell type (for the remaining 50 ct-SVGs). For each simulation scenario, we set a baseline simulation setting with $\theta=0.95$, $f = 2$, $prop = 0.2$ and $n = 5,000$. We then varied each of the four parameters

(θ, f, pi, n) one at a time on top of the baseline setting to create a total of eight simulation settings per scenario, with five simulation replicates per setting.

For the gradient pattern, we simulated either an increasing (for 50 ct-SVGs) or a decreasing (for the remaining 50 ct-SVGs) spatial gene expression pattern along the y-axis, for a fixed proportion of randomly selected cells (pi) from the target cell type. We set pi to be either 0.1, 0.15 or 0.2, corresponding to low, moderate, or high ct-SVGs signal strength, respectively. To do so, for the selected cells, we first simulated their gene expression from a negative binomial distribution and then ordered their expression levels based on their y-axis coordinates in an increasing or decreasing order. For each simulation scenario, we set a baseline simulation setting with $\theta=0.95$, $pi = 0.15$ and $n = 5000$. We then varied each of the three parameters (θ, pi, n) one at a time on top of the baseline setting to create a total of seven simulation settings per scenario, with five simulation replicates per setting.

Finally, for simulating spot-resolution spatial transcriptomics data, we followed[75] to first simulate single-cell resolution spatial transcriptomics data as described above, created a set of square grids on the tissue, treated each square grid as a spot, aggregated the expression counts of single cells that reside within the grid to serve as the spot-resolution expression, and retained spots with more than 2% of the genes containing non-zero expression level. In the simulations, we created a 70 (x-axis) x 71 (y-axis) square grid to obtain a total of 4970 spots. To mimic a spot-resolution spatial transcriptomics data such as 10X Visium, which typically contains 1–10 cells per spot, we generated 20,000 cells to ensure that each spot consists of 4-5 cells. We also obtained the actual cell type compositions on each spot. In addition, we used CARD or RCTD to estimate the cell type compositions on each spot, with the simulated single-cell resolution spatial transcriptomics data serving as the reference. We used either the actual or the estimated cell type compositions as input for ct-SVG analysis.

Additionally, we performed simulations to examine the influence of coordinates scaling on the performance of Celina, in terms of both type I error control and power. Specifically, we randomly generated spatial locations with x and y coordinates both drawn from random-point Poisson process. Both coordinates range from 0 to 1. We then multiplied the x-coordinates by a factor of 2 to create a rectangular-shaped tissue section. Afterwards, we allocated cells to these locations and simulated the spatial transcriptomics data in the same way described above. We applied two versions of Celina to analyze the data: the unscaled version of Celina that is applied to analyze the rectangular-shaped tissue section; and the scaled/default version of Celina that first scaled the tissue coordinates, thus effectively applying to analyze the distorted, square-shaped tissue section.

Compared methods

Besides Celina, we adapted and examined four existing methods originally developed for other analytic settings towards ct-SVG analysis in both simulations and real data applications. These methods include SPARK[8], SPARK-X[10], CSIDE[18], and CTSV[16], which are all applied with default settings. We did not apply CTSV in the real datasets because of its heavy computational burden.

We adapted SPARK and SPARK-X for ct-SVG analysis. Such adaptation depends on the resolution of the spatial transcriptomics data. Specifically, in single-cell resolution spatial transcriptomics data, we examined one cell type at a time and performed two analyses to detect ct-SVGs. In the first analysis, we directly applied these methods to the extracted cells from the specific cell type to detect ct-SVGs. We call this first analysis SPARK_extract or SPARK-X_extract. In the second analysis, we followed[10] and created a set of binary indicator variables, one for each cell type, to indicate whether a cell belongs to a particular cell type (1) or not (0). We then included these indicator variables as covariates in the SPARK-X model to detect ct-SVGs. We call this second

analysis SPARK-X_control. We were unable to apply SPARK for the second analysis due to its heavy computational burden. In spot-resolution spatial transcriptomics data, we also performed two analyses to detect ct-SVGs. In the first analysis, we applied these methods on all spots without adjusting for cell type composition. We call this first analysis SPARK or SPARK-X. For the second analysis, we applied SPARK-X on all spots while adjusting for true cell type composition on the spots as covariates. We call this analysis SPARK-X_control. Because SPARK-X_control had low power even with true cell type composition, we did not explore the use of estimated cell type composition from CARD or RCTD as covariates in SPARK-X. We also did not apply SPARK for the second analysis due to its heavy computational burden.

We adapted CSIDE by applying the nonparametric version of CSIDE to detect ct-SVGs through three analyses. In the first analysis, we used the "*import_weights*" function to directly incorporate the true cell type labels (for single-cell resolution data) or the true cell type composition matrix (for spot-resolution data). Afterwards, we used "*run.CSIDE.nonparam*" to detect ct-SVGs, where the "*doublet_mode*" parameter was set to "F" to incorporate the true cell type labels or the true cell type composition matrix during model fitting following the guidance provided on GitHub. One essential step in the CSIDE software is performing cell type filtering, which partitions the tissue into four equal parts and retains only the cell types that reside in every part. This filtering step in CSIDE unfortunately rendered it unsuitable for analyzing multiple simulation settings. Indeed, CSIDE failed to yield any results in 23 out of the 92 settings for single-cell resolution alternative simulations and 17 out of the 68 settings for spot-resolution alternative simulations. Manually removing the filtering step in CSIDE, unfortunately, leads to a significant reduction in the number of analyzable genes. Specifically, an average of 53% of the total genes were devoid due to this modification across all settings in scenario I. Consequently, we opted to retain the default configurations for CSIDE and did not report the results for the 23 and 17 failed settings for the single cell and spot resolution alternative simulations, respectively. Because we directly used the true cell type composition (which is unknown in real datasets), we call this analysis CSIDE (for single-cell resolution data) or CSIDE_oracle (for spot-resolution data). In the second analysis, we used the estimated cell type proportion output from its default paired method RCTD (CSIDE_wRCTD) for the simulated spot-resolution spatial transcriptomics data. In the third analysis, we used the estimated cell type proportion output from CARD (CSIDE_wCARD). While the default setting of CSIDE utilizes 15 basis functions to construct spline functions, we explored additional variants of CSIDE by reducing to either 5 (CSIDE_basis_5) or 10 (CSIDE_basis_10) basis functions.

CTSV directly models the count data to identify cell type-specific spatially variable genes for spatial transcriptomics data. CTSV tests for each gene whether its cell type-specific expression in a target cell type is associated with either the horizontal (x-axis) or the vertical (y-axis) direction. Therefore, we used two versions of CTSV, denoted as CTSV_x and CTSV_y, to identify genes with spatial expression variation in either direction. We compared to CTSV only in the main simulation settings and in one of the real datasets due to its heavy computational burden.

For Celina, for spot-resolution data, in addition to supplying the true cell type composition matrix (Celina_oracle), we also paired Celina with the estimated cell type composition estimates from either CARD (Celina_wCARD) or RCTD (Celina_wRCTD).

Analyzed datasets

We applied Celina and the other methods to analyze five public spatial transcriptomics datasets described below. We only applied SPARK or SPARK-X to the last single-cell resolution Axolotl brain data, because these two methods displayed heavily inflated type I error control in the other datasets as consistent with the spot-resolution simulations.

Human lung cancer data by 10X Visium. We obtained the 10X Visium data on human lung cancer from the 10X genomics public datasets (https://www.10xgenomics.com/resources/datasets/human-lung-cancer-ffpe-2-standard). The raw data contains 18,066 genes on 3858 spatial locations. We removed mitochondrial genes and retained genes with non-zero expression level on at least 20 locations. We also retained locations with non-zero expression for at least 20 genes. These filtering steps lead to a set of 17,257 genes on 3813 locations for analysis. We focused on 24 cell types that also appeared in a reference single cell data and performed cell type deconvolution analysis (described below) to obtain the cell type composition estimates on 3813 locations following the default QC steps in CARD. Among the 24 cell types, we further filtered out 10 minor cell types with cell type proportion below 5% across locations. We then followed the default cell type QC steps in CSIDE to focus on the remaining 14 cell types (NK cell, macrophage, endothelial cell, T cell CD4, B cell, tumor cells, T cell regulatory, stromal, plasma cell, T cell CD8, macrophage alveolar, transitional club/AT2, alveolar cell type 1, and ciliated cells). For each of the 14 cell types in turn, we followed the default QC steps in CSIDE to filter out lowly expressed genes whose total average expression counts in the cell type of focus is below 15 UMIs in the single cell reference dataset, as well as genes with expression level smaller than half of the maximum expression in the cell type of interest within the single cell reference dataset. We analyzed 369–2864 genes per cell type across the 14 cell types.

Human kidney cancer data by 10X Visium. We obtained the 10X Visium data in tumor core and tumor interface on human kidney cancer collected on the same patient (ID PD47171) from[53] (https://data.mendeley.com/datasets/g67bkbnhhg/1). The raw data of tumor core contains 36,601 genes measured on 3008 spatial locations. We removed mitochondrial genes and retained genes with non-zero expression level on at least 20 locations. We also retained locations with non-zero expression for at least 20 genes. These filtering steps led to a set of 14,837 genes measured on 2917 locations for analysis. We focused on 12 cell types that also appeared in a reference single cell data and performed cell type deconvolution analysis (described below) to obtain cell type composition estimates on 2917 locations following the default QC steps in CARD. The raw data of tumor interface contains 36,601 genes measured on 2048 spatial locations. We removed mitochondrial genes and retained genes with non-zero expression level on at least 20 locations. We also retained locations with non-zero expression for at least 20 genes. These filtering steps lead to a set of 15,413 genes on 2048 locations for analysis. We focused on 12 cell types that also appeared in a reference single cell data and performed cell type deconvolution analysis (described below) to obtain cell type composition estimates on 2048 locations following the default QC steps in CARD. In both datasets, among the 12 cell types, we further filtered out 2 minor cell types with a cell type proportion below 1% across all locations. We then followed the default cell type QC steps in CSIDE to focus on the remaining 10 cell types that include B cell, plasma cell, endothelial cell (EC), renal cell carcinoma cell (RCC), non-proximal tubule epithelial cell (Epi_non-PT), proximal tubule epithelial cell (Epi-PT), fibroblast cell, myeloid cell, plasmacytoid dendritic cell (pDC), and mast cell. For each of the 10 cell types in turn, we followed the default QC steps in CSIDE to filter out lowly expressed genes whose total average expression counts in the cell type of focus is below 15 UMIs in the single cell reference dataset, as well as genes with expression level smaller than half of the maximum expression in the cell type of interest in the single cell reference dataset. We analyzed a total of 531–1454 genes per cell type across the 10 cell types in tumor core data, and 530–1484 genes per cell type in tumor interface data.

Mouse hippocampus data by STARmap-PLUS. We obtained the STARmap-PLUS data on mouse hippocampus from the Broad Single

Cell Portal (ID SCP1375; data id: spatial_13months-disease-replicate_1)[87], which contains 2766 genes measured on 10,372 segmented cells. We retained genes with non-zero expression level on at least 20 locations and retained locations with non-zero expression for at least 20 genes. These filtering steps lead to a set of 2765 genes on 10,372 cells for analysis. We focused on 27 cell types that also appear in a reference single cell data and performed cell type deconvolution analysis (described below) to obtain the cell type composition estimates on 10,372 cells following the default QC steps in CARD. Among the 27 cell types, we further filtered out 4 minor cell types with an average cell type proportion below 1% across all cells and further followed the default cell type QC steps in CSIDE to focus on 23 cell types (interneuron, neuron, entorhinal cortex, deep layer subiculum, entorhinal cortex (IEG), subiculum, dentate principal cells, medial entorhinal cortex, CA1 Principal cells, medial entorhinal cortexm, postsubiculum, CA2 Principal cells, CA3 Principal cells, astrocyte, oligodendrocyte, myelinProcesses, polydendrocyte, microglia, resident macrophage, ependymal, endothelial stalk, mural, and choroid_plexus). For each of the 23 cell types in turn, we followed the default QC steps in CSIDE to filter out lowly expressed genes whose total average expression counts in the cell type of focus is below 15 UMIs in the single cell reference dataset, as well as genes with expression level smaller than half of the maximum expression in the cell type of interest within the single cell reference dataset. We analyzed a total of 98–621 genes per cell type across the 23 cell types.

Axolotl brain data by Stereo-Seq. We obtained the Stereo-Seq data on axolotl brain at stage 57 from https://db.cngb.org/stomics/artista/download/[4]. The raw data contains 22,893 genes on 4410 spatial locations. We removed mitochondrial genes and retained genes with non-zero expression level on at least 20 locations. We also retained locations with non-zero expression for at least 20 genes. These filtering steps lead to a set of 16,508 genes on 4410 locations for analysis. We focused on six cell types emphasized in the original study that include CP (choroid plexus, $n = 382$), EGC (ependymoglial cells, $n = 1190$), Immature Neuron ($n = 1626$), Mature Neuron ($n = 201$), NBL (neuroblasts, $n = 606$), and VLMC (vascular leptomeningeal cell, $n = 405$). For each of the six cell types in turn, we followed the default QC steps in CSIDE to filter out lowly expressed genes whose total average expression counts in the cell type of focus is below 15 UMIs as well as genes with expression level smaller than half of the maximum expression in the cell type of interest. We analyzed a total of 2424–3188 genes per cell type across the six cell types.

Detailed analyses

Data normalization. For all five datasets, we applied the *normalizeCounts* function in the *scater* R package[29] to obtain normalized gene-expression levels for analysis. This normalization method addresses the variability in cell density across spots by normalizing the data relative to the total counts for each spot because the total counts per spot are strongly correlated with the number of cells in each spot[88]. Consequently, this approach effectively accounts for variations in cell density across different spots, thereby ensuring consistency in the analysis.

Cell type deconvolution. We performed reference-based cell type deconvolution using CARD[71]. For the human lung cancer data, we obtained a single cell reference data on lung cancer[31] to serve as the reference for deconvolution. The reference data contains 190,199 cells in squamous cell lung carcinoma characterized in the original study. A total of 24 cell types are present in the reference squamous cell lung carcinoma cells and these include Alveolar cell type 1 ($n = 272$), Alveolar cell type 2 ($n = 863$), B cell ($n = 5378$), cDC1 ($n = 142$), cDC2 ($n = 1824$), Ciliated ($n = 594$), Club ($n = 233$), DC mature ($n = 149$),

Endothelial cell ($n = 7412$), Macrophage ($n = 7887$), Macrophage alveolar ($n = 5380$), Mast cell ($n = 1117$), Monocyte ($n = 6409$), Neutrophils ($n = 5465$), NK cell ($n = 3843$), other ($n = 1388$), pDC ($n = 635$), Plasma cell ($n = 2740$), Stromal ($n = 2269$), T cell CD4 ($n = 12,113$), T cell CD8 ($n = 12,566$), T cell regulatory ($n = 3263$), transitional club/AT2 ($n = 561$), and tumor cells ($n = 17,497$).

For the two human kidney cancer datasets in tumor core and tumor interface, we obtained a single cell reference data collected on the same patient in the original paper to serve as the reference for deconvolution. The reference data contains 36,110 cells that belong to multiple immune cell types and malignant cell types characterized in the original study. A total of 12 cell types are present in the reference and these include B cells ($n = 519$), plasma cell ($n = 144$), endothelial cell (EC, $n = 46$), renal cell carcinoma cell (RCC, $n = 2350$), non-proximal tubule epithelial cell (Epi_non-PT, $n = 21$), proximal tubule epithelial cell (Epi-PT, $n = 48$), fibroblast cell ($n = 32$), myeloid cell ($n = 5990$), plasmacytoid dendritic cell (pDC, $n = 73$), NK cell ($n = 1576$), T cell ($n = 25,307$), and mast cell ($n = 4$).

For the STARmap-PLUS mouse hippocampus data, we obtained a scRNA-seq data collected on mouse hippocampus (http://dropviz.org) to serve as the reference data for deconvolution[89]. The reference data contains 53,204 cells that belong to multiple cell types characterized in the original study. We focus our analysis on 27 major cell types that include astrocyte ($n = 7503$), CA1 Principal cells ($n = 5577$), CA2 Principal cells ($n = 330$), CA3 Principal cells ($n = 6272$), Cajal-retzius ($n = 336$), choroid plexus ($n = 22$), deep layer subiculum ($n = 509$), dentate hilum ($n = 530$), dentate Principal cells ($n = 13,265$), endothelial stalk ($n = 1,903$), endothelial tip ($n = 296$), entorhinal cortex ($n = 1580$), entorhinal cortex (immediate early genes, IEG, $n = 27$), ependymal ($n = 337$), interneuron ($n = 3035$), medial entorhinal cortex ($n = 867$), medial entorhinal cortexm ($n = 56$), microglia ($n = 467$), mural ($n = 705$), myelin processes ($n = 227$), neuron ($n = 3087$), oligodendrocyte ($n = 1969$), polydendrocyte ($n = 1046$), postsubiculum ($n = 43$), resident macrophage ($n = 78$), and subiculum ($n = 2679$).

Spatial domain analysis. In the human lung cancer dataset, we detected ten spatial domains with SpatialPCA using all ct-SVGs detected by Celina as input and compared the results with default setting in SpatialPCA using the top 3000 SVGs detected from SPARK as input. We combined the four tumor subregions detected with Celina and manually created tumor border mask based on the edge of the four tumor subregions for visualization. With the extracted spatial PCs, we also performed spatial trajectory inference on the tumor region following[75].

Clustering ct-SVGs detected by Celina. We performed clustering analysis on the ct-SVGs detected by Celina, both for the tumor cells in the human lung cancer data and for the RCC cells in the two human kidney cancer data. Specifically, we used the clustering function "*kmeans*"[90] implemented in the *stats* R package to cluster the ct-SVGs into three gene groups. We then summarized the mean expression level across the ct-SVGs in each group into a metagene for visualization.

Survival analysis. In the human lung cancer and human kidney cancer data, we performed survival analysis using the ct-SVGs identified by Celina to examine whether these genes are predictive of patient survival in the external The Cancer Genome Atlas (TCGA) data. To do so, we obtained both RNA-seq and clinical data for 482 lung cancer patients (cancer study ID: TCGA-LUSC) and 516 kidney cancer patients (cancer study ID: TCGA-KIRC) from TCGA using the web server http://gepia.cancer-pku.cn[91,92]. We categorized the patients into two groups based on low (25%) and high (75%) expression levels of each target gene following[93]. We then applied Kaplan-Meier method implemented

in the *survival* R package (v3.3-1)[94] to estimate the survival curves for the two patient groups. The *p*-values are calculated with the two-sided log-rank test to examine whether the ct-SVGs identified by Celina are predictive of patient survival.

Gene sets and functional enrichment analysis. In each spatial transcriptomics dataset, we first examined the functional importance of the ct-SVGs identified by different methods by comparing them to multiple reference gene lists from external data sources that relevant to the tissue and cell type in the spatial transcriptomics dataset.

Specifically, we obtained four gene lists for the human lung cancer data. The first gene list consists of 1652 cell type marker genes from the original study[31], 343 of which were analyzed by Celina and CSIDE. The second gene list is from the Harmonizome database and consists of lung cancer-related genes from four different datasets, including 2286 genes in DISEASES Text-mining Gene–disease Association Evidence Scores, GAD gene–disease associations, GWAS Catalog SNP-Phenotype Associations, and DISEASES Experimental Gene-Disease Association Evidence Scores. Among these genes, 932 were analyzed by Celina and CSIDE. The third gene list contains 1379 lung cancer-related genes obtained from the CancerMine database (http://bionlp.bcgsc.ca/cancermine/), 542 of which were analyzed by Celina and CSIDE. These genes are either cancer drivers, oncogenes, or tumor suppressors. The fourth gene list consists of 2862 transcription factors and their target genes obtained from[95], 1113 of which were analyzed by Celina and CSIDE.

We obtained three gene lists for the human kidney cancer data. The first gene list is from the Harmonizome database and consists of 1310 kidney cancer-related genes from four different datasets, including DISEASES Text-mining Gene–disease Association Evidence Scores, GAD gene–disease associations, GWAS Catalog SNP-Phenotype Associations, and DISEASES Experimental Gene-Disease Association Evidence Scores. Among these genes, 371 were analyzed by Celina and CSIDE. The second gene list contains 216 kidney cancer related genes obtained from the CancerMine database (http://bionlp.bcgsc.ca/cancermine/), 69 of which were analyzed by Celina and CSIDE. These genes are either cancer drivers, oncogenes, or tumor suppressors. The third gene list consists of 2862 transcription factors and their target genes obtained from[95], 732 of which were analyzed by Celina and CSIDE.

We obtained a gene list from the Harmonizome database that consists of 404 Alzheimer's disease related genes from GeneRIF Biological Term Annotations dataset, 155 of which were analyzed by Celina and CSIDE.

For each of the above gene lists in turn, we compared the number of significant ct-SVGs detected by Celina and CSIDE that overlapped with the genes in the gene list.

Besides the above gene lists, we also performed two additional functional enrichment analyses. The first analysis is gene set enrichment analyses (GSEA) performed using the *g:GOSt* function in the *gProfiler2* package. In the analysis, we used all expressed genes as background and used the default option g:SCS method in gProfiler2 for multiple testing correction. We report results at an FDR threshold of 0.05. The examined gene sets in GSEA are downloaded from the Molecular Signatures Database (MSigDB) and http://baderlab.org/GeneSets, which includes GO, Reactome, Panther, NetPath, NCI, MSigDB curated gene sets (C2 collection, excluding Reactome and KEGG), MSigDB Hallmark (H collection) and HumanCyc. The second analysis is the gene set variation analysis (GSVA). We used 50 "hallmark" pathways in the Molecular Signatures Database (MSigDB) to perform GSVA analysis on the human lung cancer dataset, where we used the normalized expression matrix as input and obtained a spot by pathway enrichment score matrix as output.

Statistics and reproducibility

All statistical calculations were implemented in R (v4.4.1; https://cran.r-project.org). The detailed statistical tests are indicated in corresponding Fig. legends and captions where applicable. No statistical method was used to predetermine sample size. The experiments were not randomized. This study does not involve group allocation that requires blinding.

Reporting summary

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

Data availability

This study made use of publicly available datasets. The 10X Visium data on human lung cancer is available at https://www.10xgenomics.com/resources/datasets/human-lung-cancer-ffpe-2-standard. The 10X Visium data in tumor core and tumor interface on human kidney cancer was collected on the same patient (ID PD47171) from[53] https://data.mendeley.com/datasets/g67bkbnhhg/1). The STARmap-PLUS data on mouse hippocampus is available at the Broad Single Cell Portal (ID SCP1375; data id: spatial_13months-disease-replicate_1)[87]. The Stereo-Seq data on axolotl brain at stage 57 is available at https://db.cngb.org/stomics/artista/download/[4]. The scRNA-seq reference data used in this study are all publicly available, including https://cellxgene.cziscience.com/collections/edb893ee-4066-4128-9aec-5eb2b03f8287 for the human lung cancer data[31], https://data.mendeley.com/datasets/g67bkbnhhg/1 for the two human kidney cancer datasets in tumor core and tumor interface from the original data paper[53], a scRNA-seq data collected on mouse hippocampus (http://dropviz.org) for the STARmap-PLUS mouse hippocampus data[89]. Source data are provided with this paper.

Code availability

The Celina software code is publicly available at https://xiangzhou.github.io/software/ and https://github.com/pekjoonwu/CELINA[96]. The source code is released under the GNU General Public License version 3 (GPL > = 3). Example codes for using Celina are publicly available at https://lulushang.org/Celina_Tutorial/index.html. All analysis codes for reproducing the results of the present study are publicly available at https://zenodo.org/records/14212866.

References

1. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
2. Moffitt, J. R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. USA* **113**, 11046–11051 (2016).
3. Moffitt, J. R. et al. High-performance multiplexed fluorescence in situ hybridization in culture and tissue with matrix imprinting and clearing. *Proc. Natl. Acad. Sci. USA* **113**, 14456–14461 (2016).
4. Wei, X. et al. Single-cell Stereo-seq reveals induced progenitor cells involved in axolotl brain regeneration. *Science* **377**, eabp9444 (2022).
5. Genomics, X. 10X Genomics: visium spatial gene expression. (2020).
6. Rodriques, S. G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
7. Stickels, R. R. et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, 313–319 (2021).
8. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).

9. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat. Methods* **15**, 343–346 (2018).

10. 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.* **22**, 184 (2021).

11. Hunter, M. V., Moncada, R., Weiss, J. M., Yanai, I. & White, R. M. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat. Commun.* **12**, 6278 (2021).

12. Wang, Q. et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer Cell* **32**, 42–56.e6 (2017).

13. Zhang, Y. et al. MetaTiME integrates single-cell gene expression to characterize the meta-components of the tumor immune microenvironment. *Nat. Commun.* **14**, 2634 (2023).

14. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet* **53**, 1334–1347 (2021).

15. Raghavan, S. et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell* **184**, 6119–6137.e26 (2021).

16. Yu, J. & Luo, X. Identification of cell-type-specific spatially variable genes accounting for excess zeros. *Bioinformatics* **38**, 4135–4144 (2022).

17. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet* **24**, 550–572 (2023).

18. Cable, D. M. et al. Cell type-specific inference of differential expression in spatial transcriptomics. *Nat. Methods* **19**, 1076–1087 (2022).

19. Kozareva, V. et al. A transcriptomic atlas of mouse cerebellar cortex comprehensively defines cell types. *Nature* **598**, 214–219 (2021).

20. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

21. Azizi, E. et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293–1308.e36 (2018).

22. Bastide, S. et al. TATTOO-seq delineates spatial and cell type-specific regulatory programs in the developing limb. *Sci. Adv.* **8**, eadd0695 (2022).

23. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).

24. Long, X., Yuan, X. & Du, J. Single-cell and spatial transcriptomics: advances in heart development and disease applications. *Comput. Struct. Biotechnol. J.* **21**, 2717–2731 (2023).

25. Ko, Y. et al. Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proc. Natl. Acad. Sci. USA* **110**, 3095–3100 (2013).

26. Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).

27. Wu, P. & Zhou, X. Statistical and computational methods for enabling the clinical and translational application of spatial transcriptomics. *Clin Transl Med.* **14**, e70119 (2024).

28. Binnewies, M. et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **24**, 541–550 (2018).

29. McCarthy, D. J., Campbell, K. R., Lun, A. T. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

30. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).

31. Salcher, S. et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503–1520.e8 (2022).

32. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database J. Biol. Databases Curation* **2016**, baw100 (2016).

33. Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R. & Jones, S. J. M. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* **16**, 505–507 (2019).

34. Jinesh, G. G. & Brohl, A. S. Classical epithelial-mesenchymal transition (EMT) and alternative cell death process-driven blebbishield metastatic-witch (BMW) pathways to cancer metastasis. *Signal Transduct. Target Ther.* **7**, 296 (2022).

35. Chapman, N. M. & Chi, H. mTOR signaling, Tregs and immune modulation. *Immunotherapy* **6**, 1295–1311 (2014).

36. Han, S. et al. Mitochondrial integrated stress response controls lung epithelial cell fate. *Nature* **620**, 890–897 (2023).

37. Jorgovanovic, D., Song, M., Wang, L. & Zhang, Y. Roles of IFN-gamma in tumor progression and regression: a review. *Biomark. Res.* **8**, 49 (2020).

38. Wang, L. et al. IFN-alpha/beta/IFN-gamma/IL-15 pathways identify GBP1-expressing tumors with an immune-responsive phenotype. *Clin. Exp. Med.* **24**, 102 (2024).

39. Wan, Q., Qu, J., Li, L. & Gao, F. Guanylate-binding protein 1 correlates with advanced tumor features, and serves as a prognostic biomarker for worse survival in lung adenocarcinoma patients. *J. Clin. Lab Anal.* **35**, e23610 (2021).

40. Zhang, F. et al. MTDH associates with m6A RNA methylation and predicts cancer response for immune checkpoint treatment. *iScience* **24**, 103102 (2021).

41. Tu, Z. et al. Pan-cancer analysis: predictive role of TAP1 in cancer prognosis and response to immunotherapy. *BMC Cancer* **23**, 133 (2023).

42. Fang, Z. et al. Low-density lipoprotein receptor-related protein 8 facilitates the proliferation and invasion of non-small cell lung cancer cells by regulating the Wnt/beta-catenin signaling pathway. *Bioengineered* **13**, 6807–6818 (2022).

43. Hung, J. J., Hsueh, C. T., Chen, K. H., Hsu, W. H. & Wu, Y. C. Clinical significance of E2F1 protein expression in non-small cell lung cancer. *Exp. Hematol. Oncol.* **1**, 18 (2012).

44. Sullivan, A. & Lu, X. ASPP: a new family of oncogenes and tumour suppressor genes. *Br. J. Cancer* **96**, 196–200 (2007).

45. Geng, J. et al. EZH2 promotes tumor progression via regulating VEGF-A/AKT signaling in non-small cell lung cancer. *Cancer Lett.* **359**, 275–287 (2015).

46. Lai, X. F. et al. Mesenchymal stromal cells attenuate alveolar type 2 cells senescence through regulating NAMPT-mediated NAD metabolism (vol 13, 12, 2022). *Stem Cell Res. Ther.* **13**, 12 (2022).

47. Lin, T. C. Updated functional roles of NAMPT in carcinogenesis and therapeutic niches. *Cancers* **14**, 2059 (2022).

48. Liu, T. T. et al. Visfatin mediates SCLC cells migration across brain endothelial cells through upregulation of CCL2. *Int. J. Mol. Sci.* **16**, 11439–11451 (2015).

49. Revel, M., Sautes-Fridman, C., Fridman, W. H. & Roumenina, L. T. C1q+ macrophages: passengers or drivers of cancer progression. *Trends Cancer* **8**, 517–526 (2022).

50. Dejana, E. & Orsenigo, F. Endothelial adherens junctions at a glance. *J. Cell Sci.* **126**, 2545–2549 (2013).

51. Bian, F. H. et al. Lung endothelial cells regulate pulmonary fibrosis through FOXF1/R-Ras signaling. *Nat. Commun.* **14**, 2560 (2023).

52. Hung, M. S. et al. Epidermal growth factor receptor mutation enhances expression of cadherin-5 in lung cancer cells. *PLoS One* **11**, e0158395 (2016).

53. Li, R. et al. Mapping single-cell transcriptomes in the intra-tumoral and associated territories of kidney cancer. *Cancer Cell* **40**, 1583–1599.e10 (2022).

54. Ellen, T. P., Ke, Q., Zhang, P. & Costa, M. NDRG1, a growth and cancer related gene: regulation of gene expression and function in normal and disease states. *Carcinogenesis* **29**, 2–8 (2008).

55. Dolezal, J. M., Dash, A. P. & Prochownik, E. V. Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* **18**, 275 (2018).

56. Schmit, K. et al. Characterization of the role of TMEM45A in cancer cell sensitivity to cisplatin. *Cell Death Dis.* **10**, 919 (2019).

57. Kang, J. et al. Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. *Signal Transduct. Target Ther.* **6**, 323 (2021).

58. Brahimi-Horn, M. C., Chiche, J. & Pouyssegur, J. Hypoxia and cancer. *J. Mol. Med.* **85**, 1301–1307 (2007).

59. Groblewska, M., Muszynski, P., Wojtulewska-Supron, A., Kulczynska-Przybik, A. & Mroczko, B. The role of visinin-like protein-1 in the pathophysiology of Alzheimer's disease. *J. Alzheimers Dis.* **47**, 17–32 (2015).

60. Deng, Y. et al. Loss of LAMP5 interneurons drives neuronal network dysfunction in Alzheimer's disease. *Acta Neuropathol.* **144**, 637–650 (2022).

61. Yoon, G. et al. Obesity-linked circular RNA circTshz2-2 regulates the neuronal cell cycle and spatial memory in the brain. *Mol. Psychiatry* **26**, 6350–6364 (2021).

62. Tan, M. G. et al. Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *J. Neurosci. Res* **88**, 1157–1169 (2010).

63. Wang, B., Liu, W. & Sun, F. Nucleosome assembly protein 1-like 5 alleviates Alzheimer's disease-like pathological characteristics in a cell model. *Front. Mol. Neurosci.* **15**, 1034766 (2022).

64. Rangaraju, S. et al. Identification and therapeutic modulation of a pro-inflammatory subset of disease-associated-microglia in Alzheimer's disease. *Mol. Neurodegener.* **13**, 24 (2018).

65. Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290.e17 (2017).

66. Ngo, C. & Kothary, R. MicroRNAs in oligodendrocyte development and remyelination. *J. Neurochem* **162**, 310–321 (2022).

67. Lu, C. D. et al. Transferrin is responsible for mediating the effects of iron ions on the regulation of anterior pharynx-derective-1α/β and Presenilin 1 expression via PGE and PGD at the early stage of Alzheimer's Disease. *Aging* **10**, 3117–3135 (2018).

68. Magnusson, J. P. et al. Activation of a neural stem cell transcriptional program in parenchymal astrocytes. *Elife* **9**, e59733 (2020).

69. Chitra, U. et al. Mapping the topography of spatial gene expression with interpretable deep learning. In *International Conference on Research in Computational Molecular Biology* pp. 368-371 (Cham: Springer Nature, Switzerland, 2024).

70. Zhao, P., Zhu, J., Ma, Y. & Zhou, X. Modeling zero inflation is not necessary for spatial transcriptomics. *Genome Biol.* **23**, 118 (2022).

71. Ma, Y. & Zhou, X. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat. Biotechnol.* **40**, 1349–1359 (2022).

72. Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* **99**, 673–686 (2004).

73. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36 (2011).

74. Wood, S. N. Thin plate regression splines. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **65**, 95–114 (2003).

75. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat. Commun.* **13**, 7203 (2022).

76. Silverman, B. W. *Density estimation for statistics and data analysis*, 175 p. (Chapman and Hall, London; New York, 1986).

77. Sheather, S. J. & Jones, M. C. A reliable data-based bandwidth selection method for Kernel density-estimation. *J. R. Stat. Soc. Ser. B Methodol.* **53**, 683–690 (1991).

78. Bartlett, M. S. Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **160**, 0268–0282 (1937).

79. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).

80. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet* **89**, 82–93 (2011).

81. Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics* **2**, 110–114 (1946).

82. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).

83. Li, Z. & Zhou, X. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biol.* **23**, 168 (2022).

84. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).

85. Zhu, J., Shang, L. & Zhou, X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biol.* **24**, 39 (2023).

86. Hoff, P. Modern applied statistics with S-PLUS. *Sociol. Methods Res.* **30**, 293–295 (2001).

87. Zeng, H. et al. Integrative in situ mapping of single-cell transcriptional states and tissue histopathology in a mouse model of Alzheimer's disease. *Nat. Neurosci.* **26**, 430–446 (2023).

88. Liu, B., Li, Y. & Zhang, L. Analysis and visualization of spatial transcriptomic data. *Front. Genet.* **12**, 785290 (2021).

89. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015 (2018).

90. Forgy, E. W. Cluster analysis of multivariate data—efficiency vs interpretability of classifications. *Biometrics* **21**, 768–76 (1965).

91. Tang, Z. et al. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98–W102 (2017).

92. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

93. Kim, N. et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).

94. Li, J. C. A. Modeling survival data: extending the Cox model. *Sociol. Methods Res.* **32**, 117–120 (2003).

95. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

96. L. Shang, X. Zhou, P.W. Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics. *pekjoonwu/CELINA: Celina_1.0.0 (v1.0.0)* https://doi.org/10.5281/zenodo.14212866 (2024).

Acknowledgements

Author contributions

X.Z. conceived the idea and provided funding support. L.S., P.W., and X.Z. designed the experiments. L.S., P.W. developed the method, implemented the software, performed simulations and analyzed real data. L.S., P.W., and X.Z. 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/s41467-025-56280-4.

Correspondence and requests for materials should be addressed to Xiang Zhou.

Peer review information *Nature Communications* thanks Xiangyu Luo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.