Spatial charting of single cell transcriptomes in tissues

Wei and He et al.

Supplementary Materials

CONTENTS

Supplementary Figures

Supplementary Figure 1 – Overview of two downstream computational modules of the CellTrek toolkit

(**a**) The SColoc module. Based on the CellTrek result map, three different spatial dissimilarity methods, i.e., KL, DT, KD, can be applied to calculate a cell-type spatial dissimilarity matrix, and an MST is used to generate a tree structure. These steps are conducted repetitively on bootstrapped samples to calculate a consensus matrix on dissimilarity matrices or MSTs, which produces a final cell-type spatial graph representation. (**b**) The SCoexp module. For cells of interest, based on the CellTrek map, SCoexp first calculates a spatial kernel matrix using RBF based on their spatial distance. Next, based on the spatial kernel matrix and cell-gene expression matrix, SCoexp calculates the spatial weighted gene co-expression. Gene modules are then identified using CC or WGCNA. For the identified co-expression modules, module activity scores can be computed and mapped back to the CellTrek coordinates.
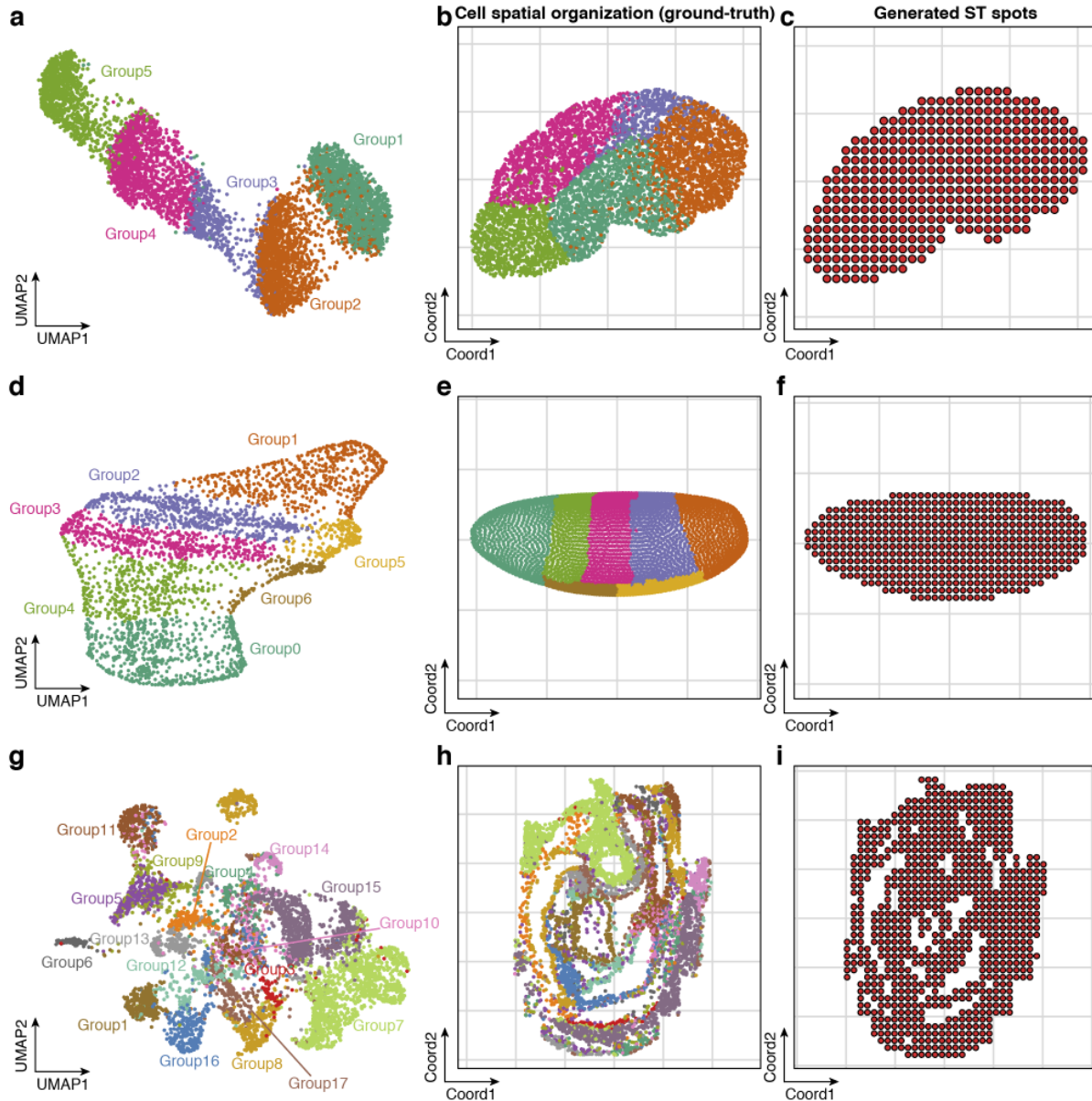
Supplementary Figure 2 – Three spatial single-cell datasets for benchmarking

(**a**) UMAP of a simulated scRNA-seq data with 5 cell groups. (**b**) Spatial organization of the simulated data as the ground truth. (**c**) ST data generated based on panel (**b**). (**d**) UMAP of *Drosophila* embryo FISH-generated single cell data. (**e**) Spatial organization of *Drosophila* embryo cells as the ground truth. (**f**) *Drosophila* embryo ST data generated based on panel (**e**). (**g**) UMAP of mouse embryo seqFISH data (Group1: Cardiomyocytes; Group2: Cranial mesoderm; Group3: Definitive endoderm; Group4: Dermomyotome; Group5: Endothelium; Group6: Erythroid; Group7: Forebrain midbrain hindbrain; Group8: Gut tube; Group9: Haematoendothelial progenitors; Group10: Intermediate mesoderm; Group11: Lateral plate mesoderm; Group12: Mixed mesenchymal mesoderm; Group13: Neural crest; Group14: Presomitic mesoderm; Group15: Spinal cord; Group16: Splanchnic mesoderm; Group17:Suface ectoderm). (**h**) Spatial organization of the mouse embryo cells as the ground truth. (**i**) mouse embryo ST data generated based on panel (**h**). In (**c**), (**f**) and (**i**), each ST spot aggregates the 5 nearest cells to generate the ST data.

4

a

Reference | CellTrek | NVSP-CellTrek | SrtCT

● Group1 ● Group2 ● Group3 ● Group4 ● Group5

b

● CellTrek ■ SrtCT ▲ NVSP-CellTrek

KL-divergence

Group1 Group2 Group3 Group4 Group5

c

Reference | CellTrek | NVSP-CellTrek | SrtCT

● Group0 ● Group1 ● Group2 ● Group3 ● Group4 ● Group5 ● Group6

d

● CellTrek ■ SrtCT ▲ NVSP-CellTrek

KL-divergence

Group0 Group1 Group2 Group3 Group4 Group5 Group6

e

eve | hb | ken | Kr | sna

level
high
low

f

Reference | CellTrek | NVSP-CellTrek | SrtCT

● Group1 ● Group3 ● Group5 ● Group7 ● Group9 ● Group11 ● Group13 ● Group15 ● Group17
● Group2 ● Group4 ● Group6 ● Group8 ● Group10 ● Group12 ● Group14 ● Group16

g

● CellTrek ■ SrtCT ▲ NVSP-CellTrek

KL-divergence

Group11 Group7 Group6 Group8 Group15 Group16 Group3 Group2 Group12 Group10 Group5 Group9

h

Gut tube | Foxa1 | Cldn4

level
high
low

i

Comp. 2

Comp. 1

Pseudotime

Pseudotime

5

Supplementary Figure 3 – Benchmarking on three datasets with known spatial references
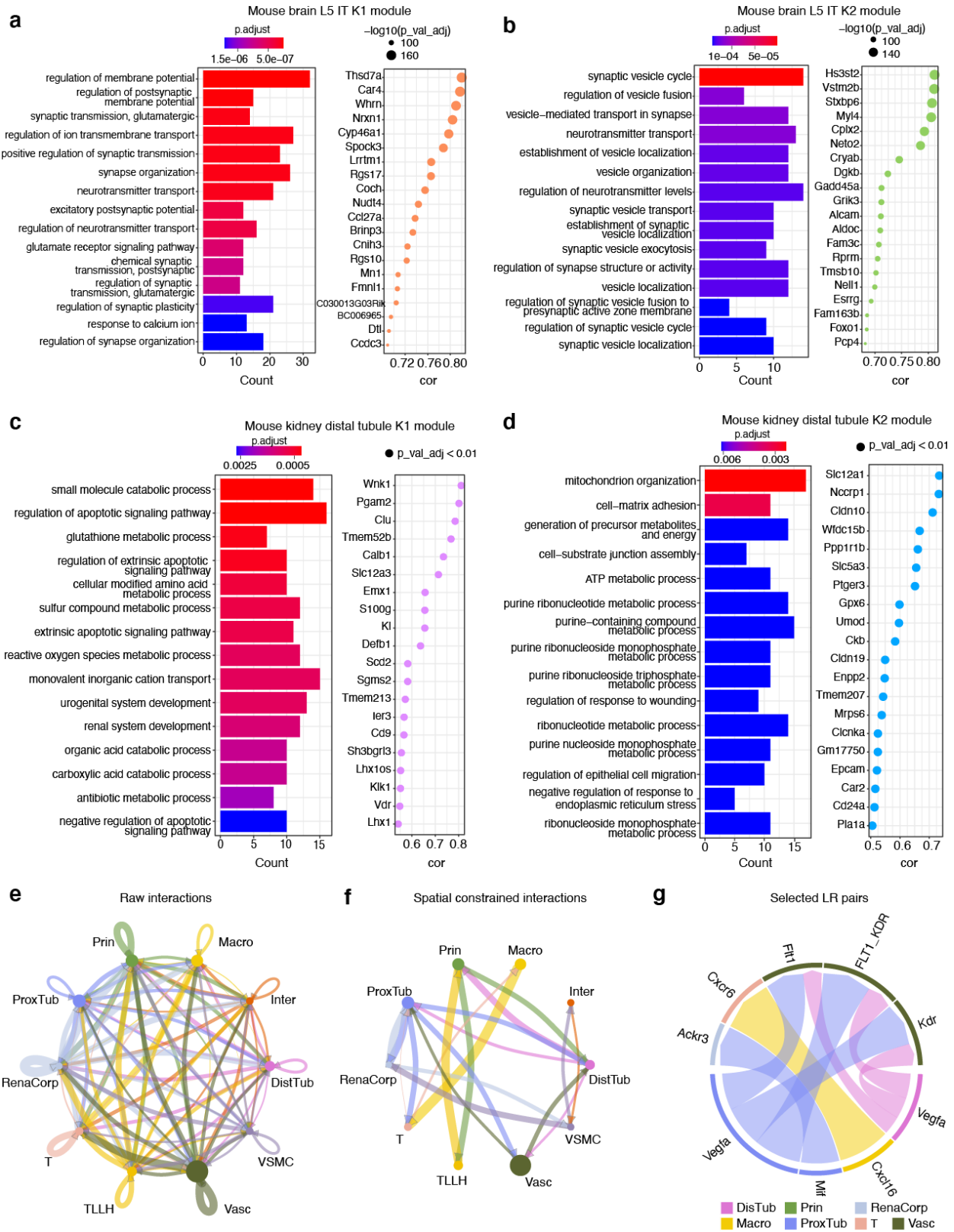
(**a**) Comparison of the ground-truth reference, CellTrek, NVSP-CellTrek and SrtCT results on the simulated data. (**b**) KL-divergence of spatial cell charting methods for each cell type compared with the ground-truth. (**c**) Comparison of the ground-truth reference, CellTrek, NVSP-CellTrek and SrtCT results on the *Drosophila* embryo data. (**d**) KL-divergence of spatial cell charting methods for each cell type compared with the ground-truth. (**e**) The expression of *Drosophola* embryogenic marker genes (*eve*, *hb*, *ken*, *Kr* and *sna*) on the CellTrek map. (**f**) Comparison of the ground-truth reference, CellTrek, NVSP-CellTrek and SrtCT results on the mouse embryo data. (**g**) KL-divergence of three spatial cell charting methods for each cell type compared with the ground-truth. (**h**) The CellTrek map of a subgroup of gut tube (foregut cells) and the marker genes (*Foxa1* and *Cldn4*). (**i**) Trajectory analysis for foregut tube cells (left) and CellTrek spatial mapping of the pseudotime values in the tissue section (right).
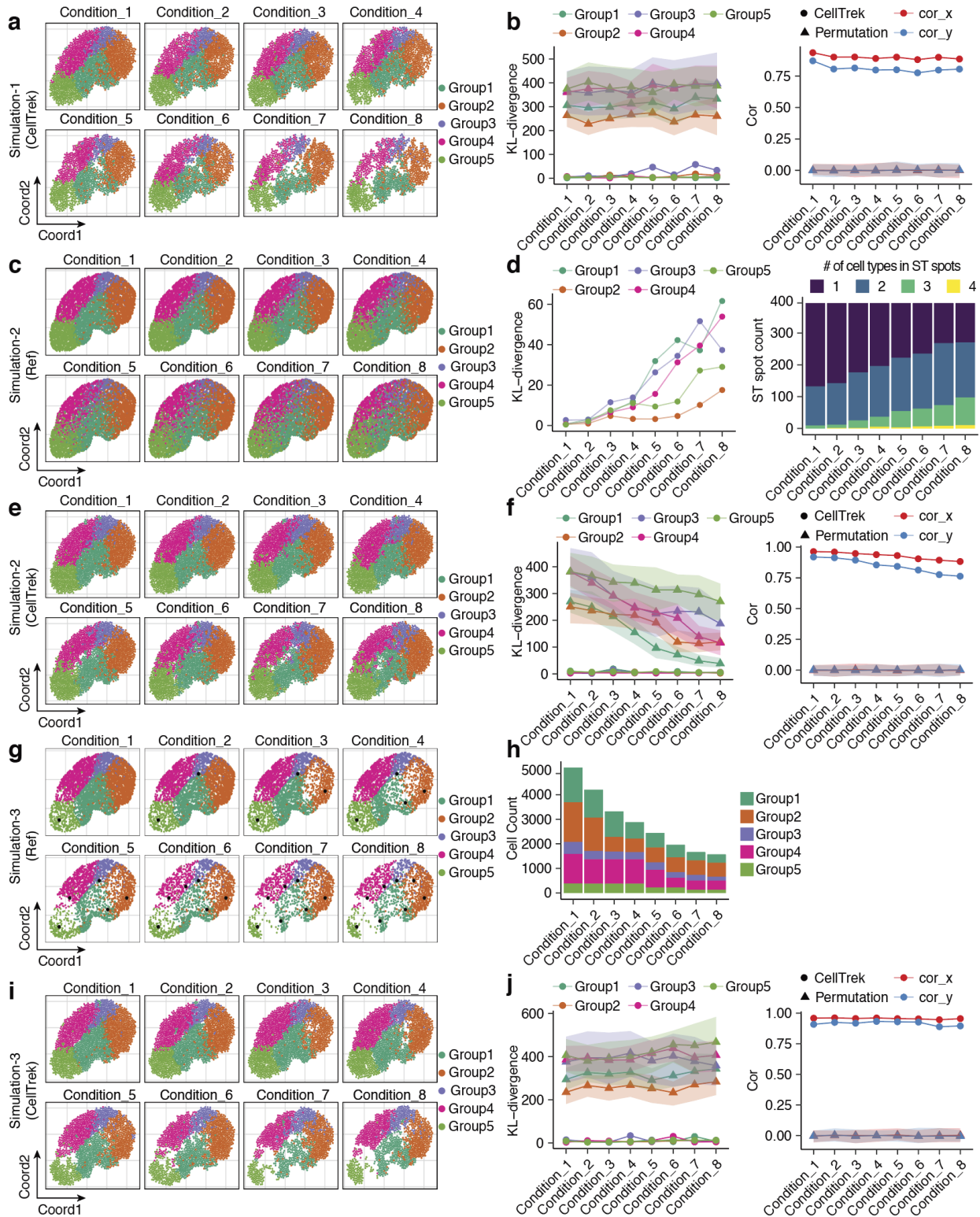
**Supplementary Figure 4 – Simulation study of CellTrek under different data conditions**

(**a**) CellTrek results under different gene counts in the simulated data. From condition1 to condition8, the gene counts decreased. (**b**) KL-divergence (left) and spatial coordinate

correlation (right) of CellTrek and permutation test across different gene count conditions. (**c**) Simulation of the increasing spatial randomness of the reference data from condition1 to condition8. (**d**) KL-divergence of the simulated reference compared with the raw spatial cell organization (left) and the number of cell types in ST spot (right) across different spatial randomness conditions. (**e**) CellTrek results under different spatial randomness of the reference data. (**f**) KL-divergence (left) and spatial coordinate correlation (right) of CellTrek and permutation test across different spatial randomness conditions. (**g**) Simulation of the decreasing tissue spatial densities (cell numbers) of the reference data from condition1 to condition8. Black dots display the marker points for spatial down-sampling. (**h**) Bar plot showing the number of cells under different tissue spatial density conditions. (**i**) CellTrek results under different tissue densities of the reference data. (**j**) KL-divergence (left) and spatial coordinate correlation (right) of CellTrek and permutation test across different tissue spatial density conditions. Detailed parameter settings are provided in Supplementary Table 1-3. For the permutation tests, the ribbon shows the 95% confidence intervals.

a Mouse brain L5 IT K1 module

b Mouse brain L5 IT K2 module

c Mouse kidney distal tubule K1 module

d Mouse kidney distal tubule K2 module

e Raw interactions

f Spatial constrained interactions

g Selected LR pairs

9

Supplementary Figure 5 – Mouse brain and kidney co-expression module enrichment analysis and mouse kidney spatial constrained cell-cell interaction analysis

(**a-b**) GO enrichment analyses (left) and module-correlated genes (right) for mouse brain L5 IT K1 and K2 modules, respectively. (**c-d**) GO enrichment analyses (left) and module-correlated genes (right) for mouse kidney distal tubule K1 and K2 modules, respectively. (**e**) Summary of cell-cell interaction network of mouse kidney cells using CellChat. (**f**) Spatial constrained cell-cell interaction network of mouse kidney cells. (**g**) Chord plot showing some identified ligand-receptor pairs. The edge width in (**e**) and (**f**) represent the total interaction strength; and the arc width in (**g**) represent the interaction score calculated from CellChat. The direction of the arrows goes from the ligands to the receptors. In **a-d**, *P* values were calculated by hypergeometric distribution and adjusted by Benjamini-Hochberg method. Module-correlated genes were determined by the CellTrek *FindCorMarkers* module. Spearman correlation was performed and *P* values were adjusted by the Bonferroni method.
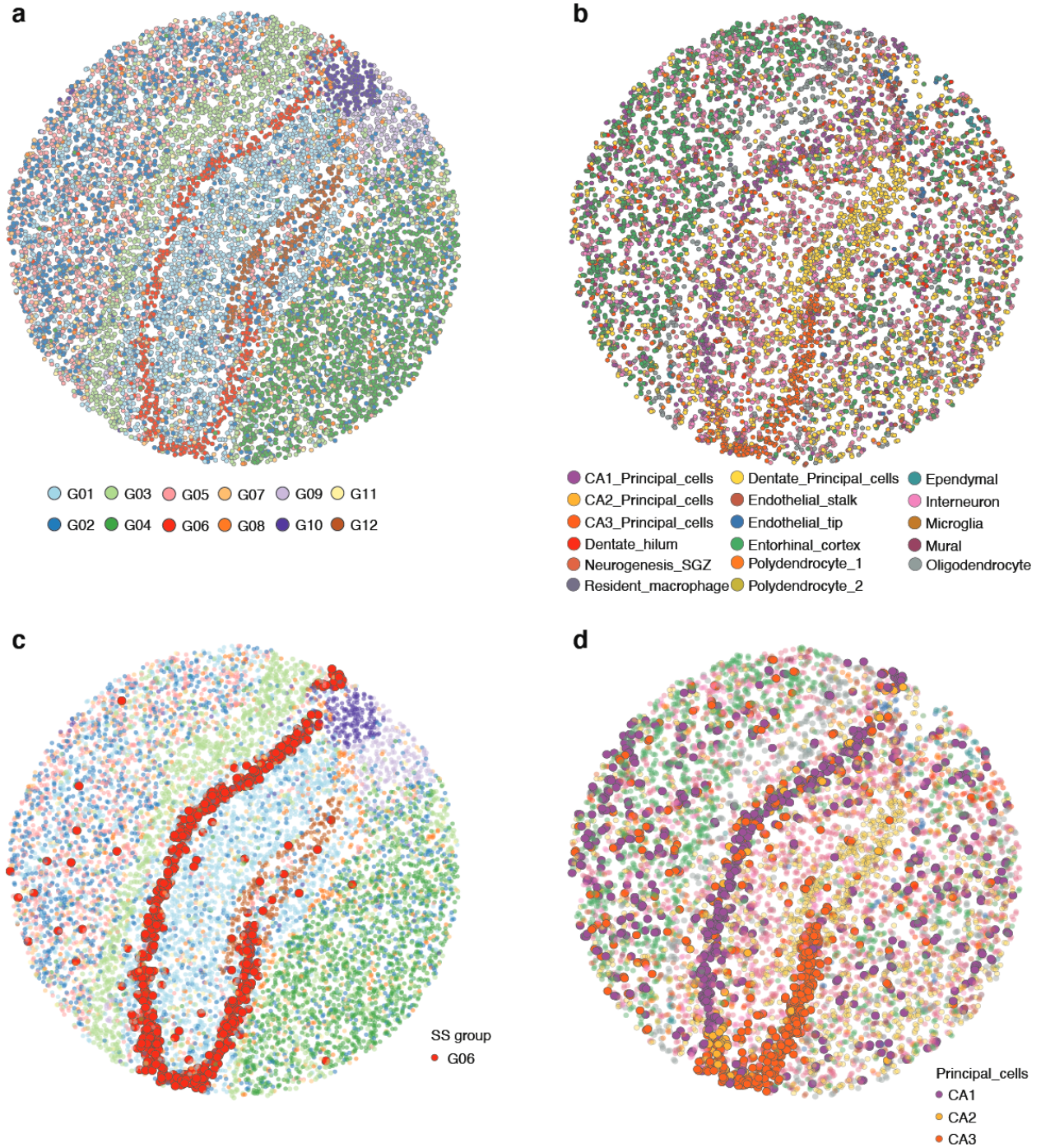
Supplementary Figure 6 – CellTrek analysis of mouse hippocampus Slide-seq data
(**a**) Unsupervised clustering of mouse hippocampus Slide-seq data identified 12 clusters. (**b**) CellTrek spatial mapping of scRNA-seq data to the tissue section. (**c**) Slide-seq data with G06 group highlighted. (**d**) CellTrek map with three principal cells (CA1, CA2 and CA3) highlighted which corresponded to G06.

Supplementary Figure 7 – ScRNA-seq and ST analyses of DCIS1

(**a**) UMAP of DCIS1 scRNA-seq data. (**b**) Consensus heatmap of scRNA-seq data for top differentially expressed genes for each cell type. (**c**) CellTrek result of three tumor subclones (clone1-3) which is also presented in **Fig. 4e**. (**d**) Unsupervised clustering of ST tumor spots identified five clusters (ST1-5). (**e**) Gene expression-based Spearman correlation analysis between three tumor subclones in the scRNA-seq data (clone1-3) and five ST clusters (ST1-5).
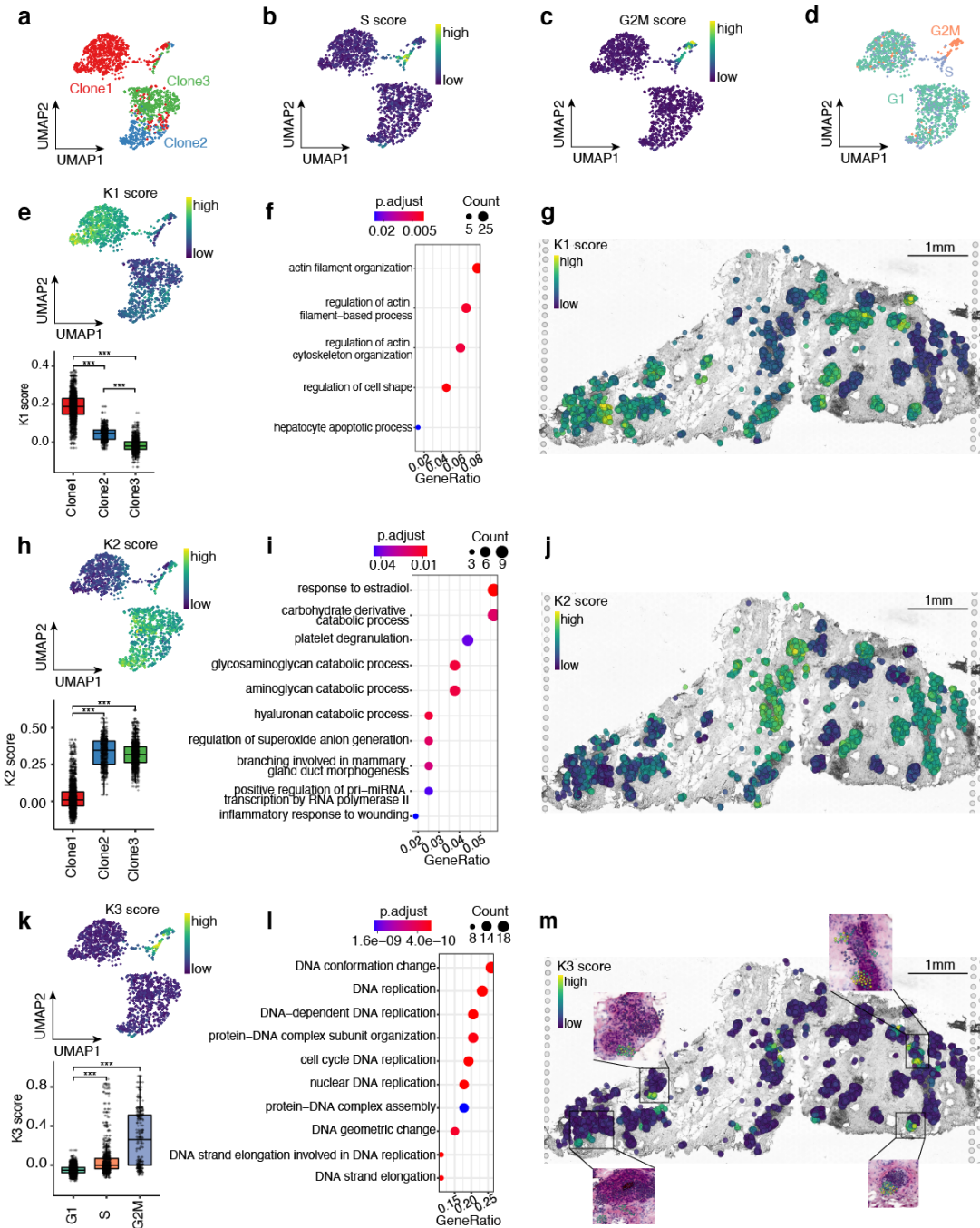
Supplementary Figure 8 – Spatial co-expression patterns of tumor cells in DCIS1

(**a**) Transcriptome-based UMAP of three tumor subclones identified from DCIS1 scRNA-seq data (Clone1 n = 1,295; Clone2 n = 523; Clone3 n = 862). Seurat-based cell cycle analysis on tumor cells showed S scores (**b**), G2M scores (**c**) and categorical cell cycle phases (G1 n = 2,060; G2M n = 508; S n = 1,019) (**d**). (**e**) K1 module scores on the UMAP (upper) and boxplot across tumor subclones (lower). (**f**) GO enrichment analysis for the K1 module. (**g**) CellTrek map of K1 module scores. (**h**) K2 module scores on the UMAP (upper) and boxplot across tumor subclones (lower). (**i**) GO enrichment analysis for the K2 module. (**j**) CellTrek map of K2 module scores. (**k**) K3 module scores on the UMAP (upper) and boxplot across different cell cycle phases (lower). (**l**) GO enrichment analysis for the K3 module. (**m**) CellTrek map of K3

13

module scores. In e, h, and k two-sided Wilcoxon rank-sum tests were performed. Boxplots show the median with interquartile ranges (25–75%); whiskers extend to 1.5X the interquartile range from the box. In **f**, **i** and **l**, *P* values were calculated by hypergeometric distribution and further adjusted by Benjamini-Hochberg method.

Supplementary Figure 9 – ScRNA-seq data analysis of DCIS2

(**a**) UMAP of DCIS2 scRNA-seq data, with clusters corresponding to different cell types. (**b**) Consensus heatmap of top differentially expressed genes for each cell cluster. (**c**) Heatmap of CopyKAT inferred copy number profiles from the scRNA-seq data and hierarchical clustering of cells. (**d**) Violin plots showing expression of canonical T cell state markers across different T cell states. (**e**) Consensus heatmap showing expression of top differentially expressed genes across different myeloid cell states.
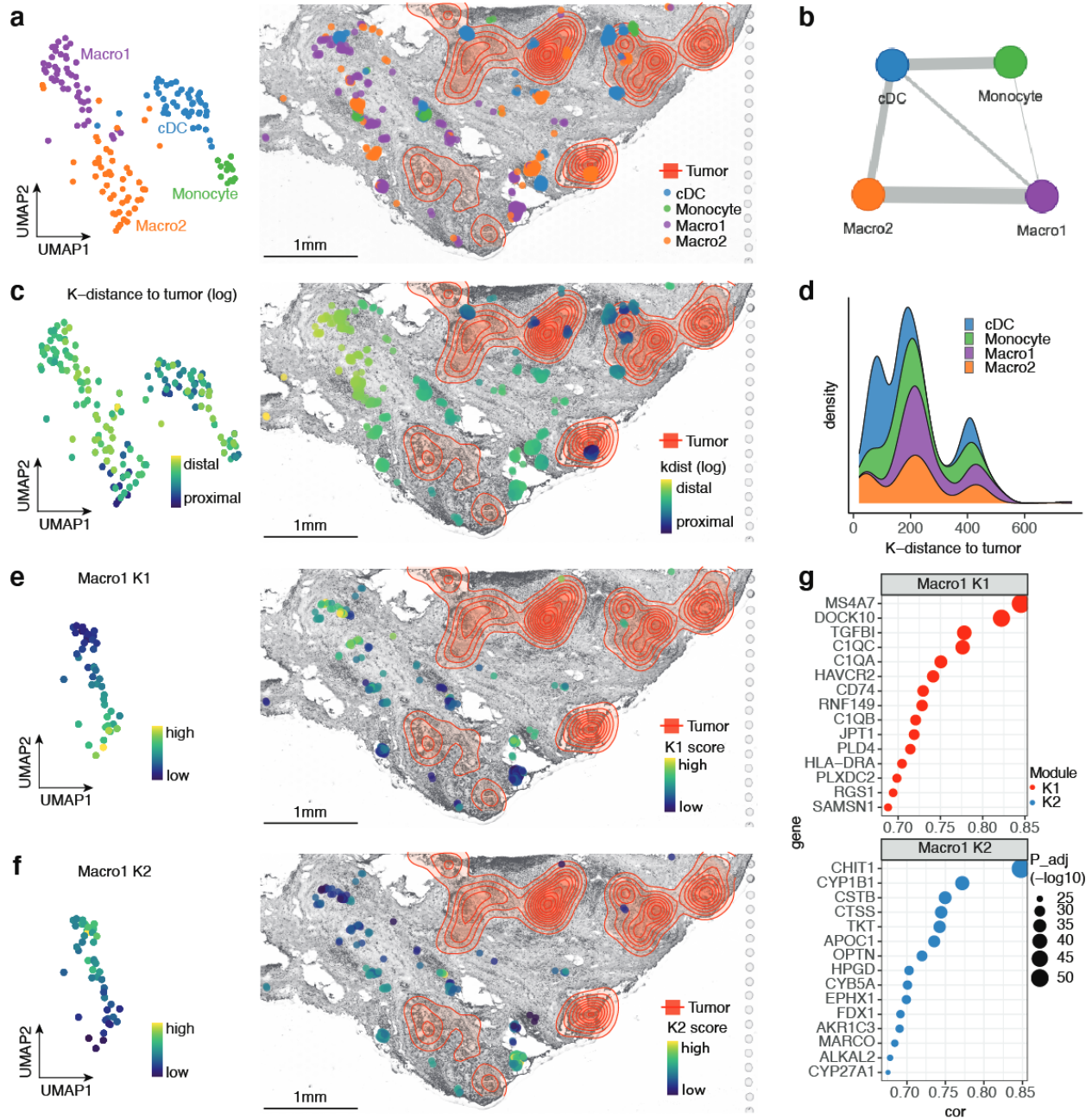
Supplementary Figure 10 – CellTrek analyses of myeloid cell states in DCIS2

(**a**) UMAP representation of myeloid cell states (left) and corresponding CellTrek map (right) (**b**) Spatial graph of myeloid cell states colocalization using SColoc. (**c**) UMAP (left) and CellTrek map (right) of myeloid cells K-distance to tumor cells. (**d**) Density plot of K-distances to tumor cells for each myeloid cell state. (**e**) UMAP (left) and CellTrek map (right) of Macro1 K1 module scores. (**f**) UMAP (left) and CellTrek map (right) of Macro1 K2 module scores. (**g**) Macro1 K1 (top) and K2 (bottom) module top correlated genes. The contour plots represent the densities of tumor cells. Module-correlated genes were determined by the CellTrek *FindCorMarkers* module. Spearman correlation was performed and *P* values were adjusted by the Bonferroni method.
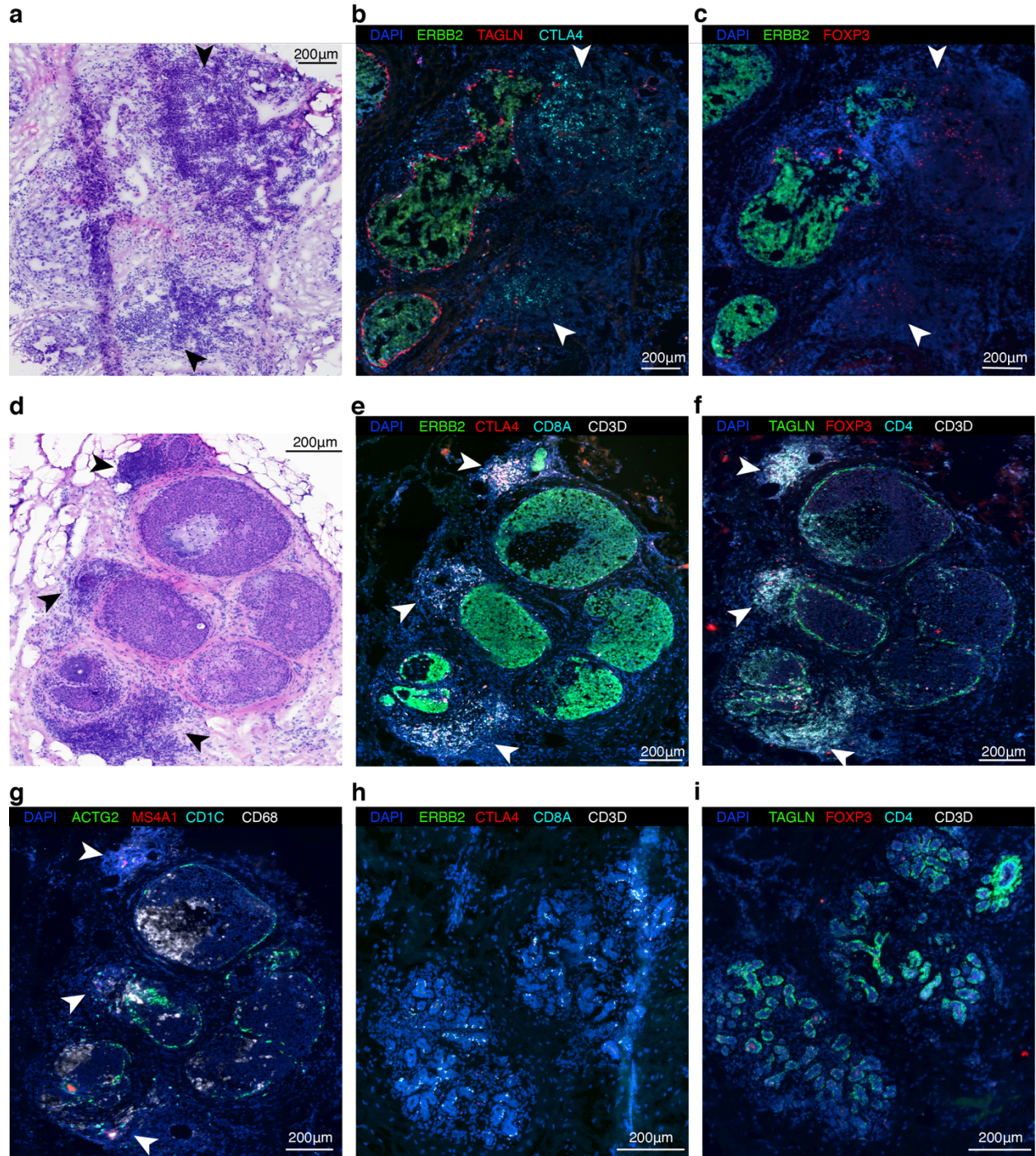
Supplementary Figure 11 – Orthogonal validation of tumor and immune cell spatial distributions using RNAscope assay in DCIS2 and DCIS3

(a) H&E image of the DCIS2 tumor region. (b) RNAscope assay for *ERBB2, TAGLN* and *CTLA4* in DCIS2 tumor region. (c) RNAscope assay for *ERBB2* and *FOXP3* in DCIS2 tumor region. Data in (a-c) were collected from adjacent tissue slides. (d) H&E image of the DCIS3 tumor region. (e) RNAscope assay for *ERBB2, CTLA4, CD8A* and *CD3D* in DCIS3 tumor region. (f) RNAscope assay for *TAGLN*, *FOXP3*, *CD4* and *CD3D* in DCIS3 tumor region. (g) RNAscope assay for *ACTG2*, *MS4A1*, *CD1C* and *CD68* in DCIS3 tumor region. (h) RNAscope assay for *ERBB2*, *CTLA4*, *CD8A* and *CD3D* in DCIS3 adjacent normal lobular region. (i) RNAscope assay for *TAGLN*, *FOXP3*, *CD4* and *CD3D* in DCIS3 adjacent normal lobular region. Arrows point to different immune cell markers near the DCIS tumor ducts.

Supplementary Tables

Supplementary Table 1 Parameter settings for different conditions in simulation-1

Condition	scRNA-seq				ST			
	Counts	Cell Proportion	Median Count	Median Gene	Counts	Spot Proportion	Median Count	Median Gene
1	10	0.2	2815	494	50	0.2	15732	1526
2	10	0.4	2328	414	50	0.4	12516	1356
3	15	0.4	2195	414	75	0.4	12489	1356
4	15	0.5	1781	207	75	0.5	7006	338
5	20	0.5	1622	192.5	100	0.5	5055	320
6	20	0.6	929	63	100	0.6	2257	39
7	25	0.6	671	42	125	0.6	1537	23
8	25	0.7	444.5	31	125	0.7	1183	19

Counts means the UMI counts for each gene we downsampled and Cell/Spot Proportion means the proportion of cells/ST spots were used for downsampling. Median Count and Gene mean the median of UMI count and gene number after downsampling.

Supplementary Table 2 Parameter settings for different conditions in simulation-2

| Condition	Spatial Deviation
1	30
2	60
3	90
4	120
5	150
6	180
7	210
8	240

Spatial Deviation means the deviation of the spatial gaussian noise we added to the cell.

**Supplementary Table 3 Parameter settings for different marker points in simulation-3**

| MarkerPoint	X-coordinate	Y-coordinate
1	350	300
2	1300	1100
3	1950	800
4	1500	600
5	600	800
6	1000	1000
7	900	400
8	1700	1100

Supplementary Data

Supplementary Data 1 – Screenshot of the CellTrek result for mouse brain data
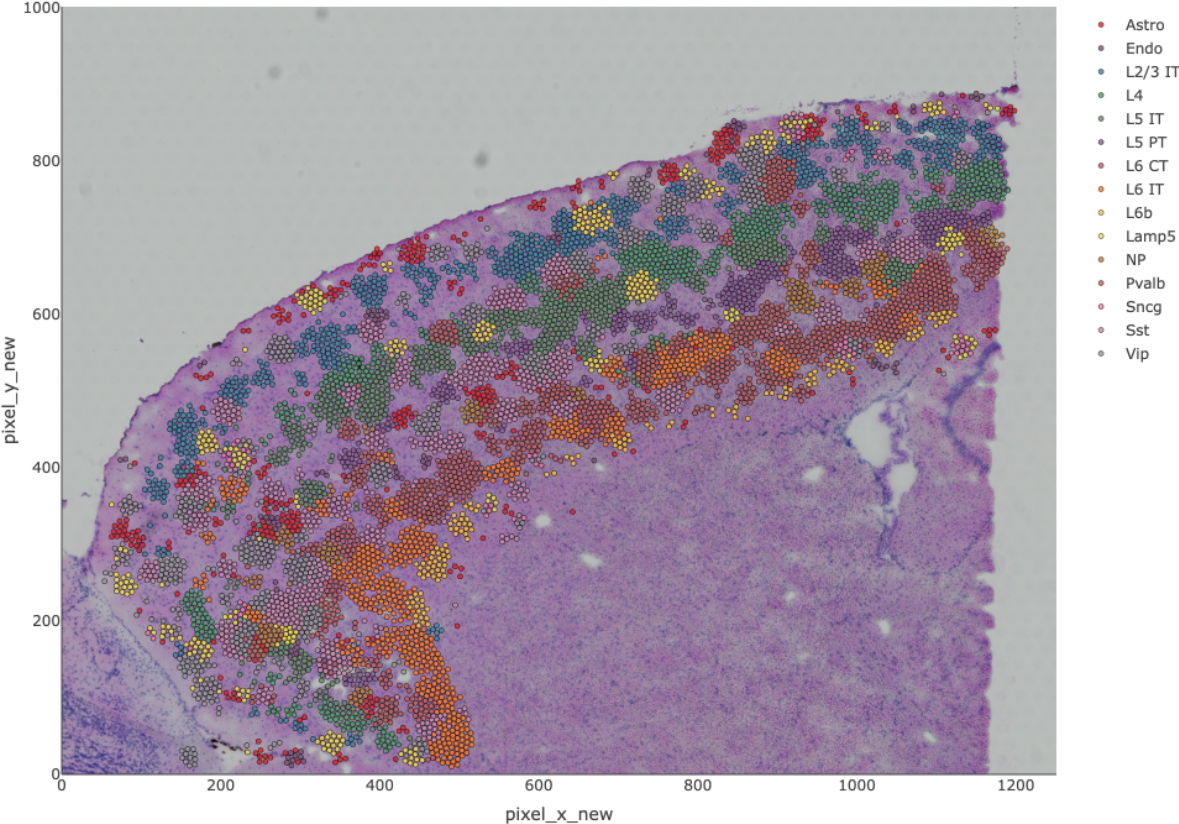
Supplementary Data 2 – Screenshot of the CellTrek result for mouse kidney data
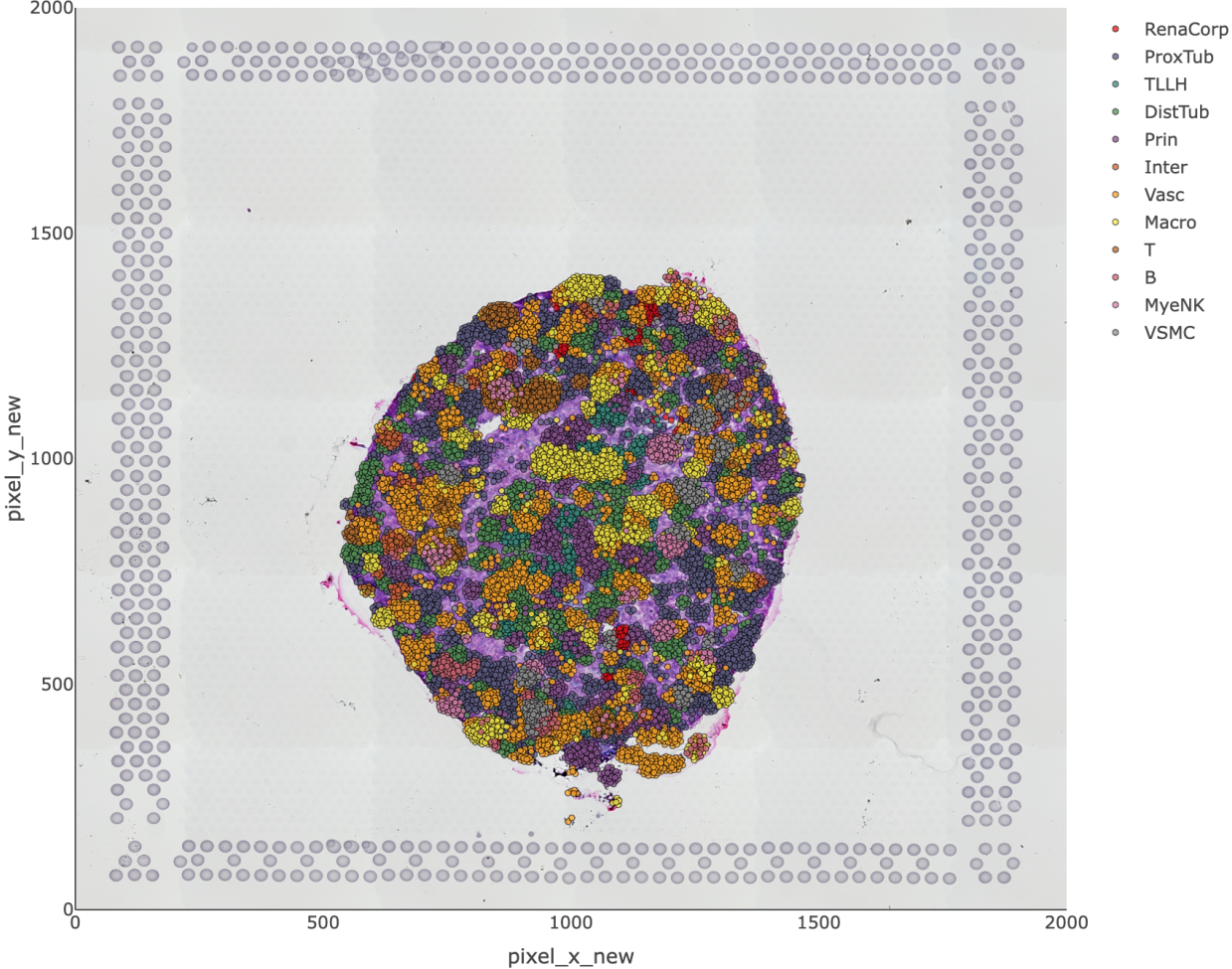
Supplementary Data 3 – Screenshot of the CellTrek result for DCIS1 data
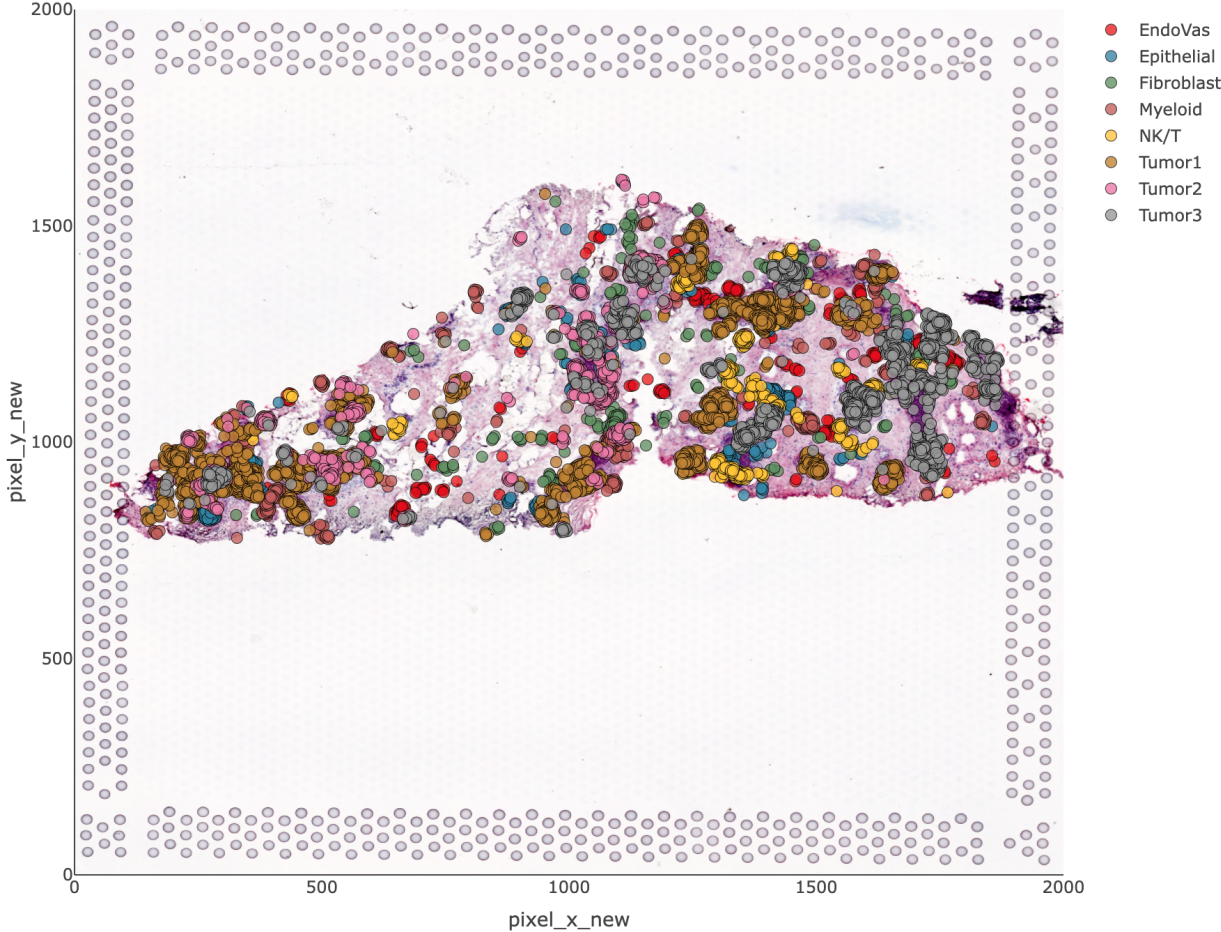
Supplementary Data 4 – Screenshot of the CellTrek result for DCIS2 data