Supplementary Information

ClusterMap for multi-scale clustering analysis of spatial gene expression

Yichun He[1,2,8], Xin Tang[1,2,8], Jiahao Huang[2], Jingyi Ren[2,3], Haowen Zhou[2], Kevin Chen[4], Albert Liu[2,3], Hailing Shi[2,3], Zuwan Lin[4,2], Qiang Li[1], Abhishek Aditham[2,5], Johain Ounadjela[2,6], Emanuelle I Grody[2,6], Jian Shu[2,6,7], Jia Liu[1*]& Xiao Wang[2,3*]

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

[2]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[3]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA.

[4]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.

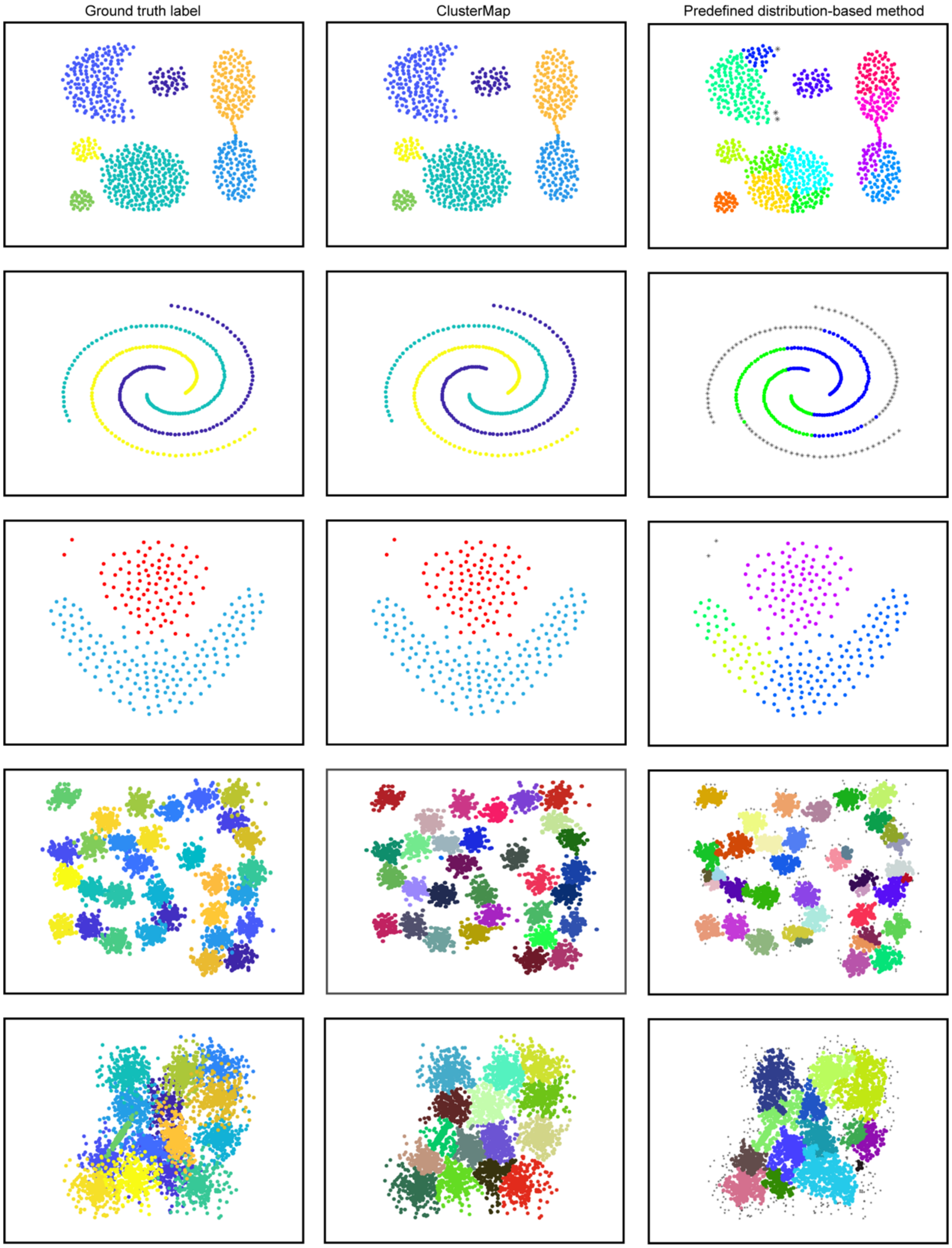
[5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

[6]Whitehead Institute for Biomedical Research, Cambridge, MA, USA.

[7]Cutaneous Biology Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

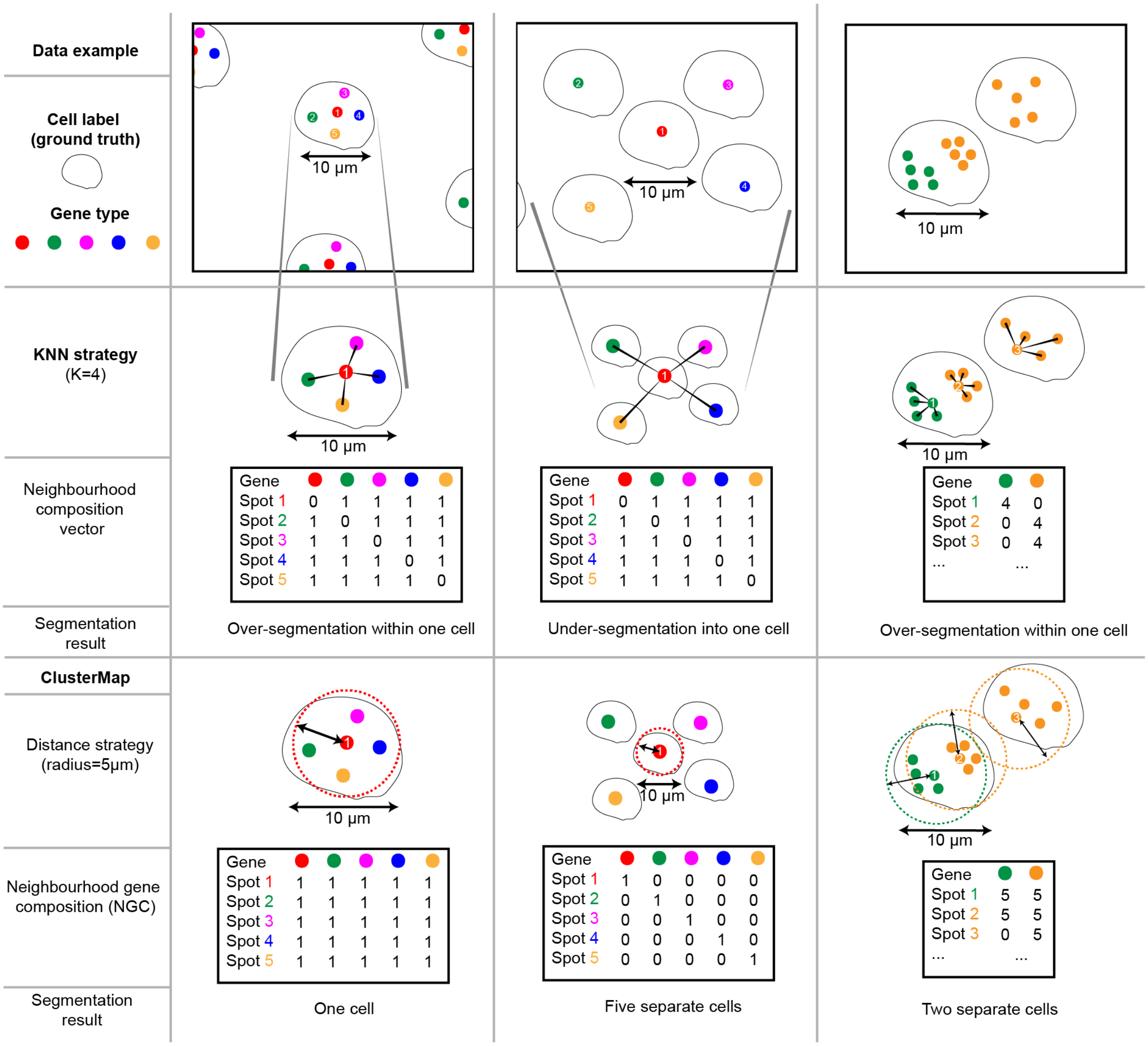
[8]These authors contributed equally: Yichun He, Xin Tang.

[*]e-mail: wangxiao@broadinstitute.org; jia_liu@seas.harvard.edu.


Ground truth label
ClusterMap
Predefined distribution-based method

Supplementary Figure 1

Performance comparison of ClusterMap and predefined distribution-based method in simulated data. Different colors represent different segmentation results. Note that the gene identity of each spot is randomly assigned from 1 to 5 as pseudo gene type. Left: ground truth; Middle: ClusterMap results; Right: results using predefined distribution-based method[20].



Data example
Cell label (ground truth)
Gene type
10 µm
1
2
3
4
5
KNN strategy
(K=4)
Neighbourhood composition vector
Segmentation result
Over-segmentation within one cell
Under-segmentation into one cell
Over-segmentation within one cell
ClusterMap
Distance strategy
(radius=5µm)
Neighbourhood gene composition (NGC)
Segmentation result
One cell
Five separate cells
Two separate cells

Gene	🔴	🟢	🟣	🔵	🟠
Spot 1	0	1	1	1	1
Spot 2	1	0	1	1	1
Spot 3	1	1	0	1	1
Spot 4	1	1	1	0	1
Spot 5	1	1	1	1	0

Gene	🔴	🟢	🟣	🔵	🟠
Spot 1	0	1	1	1	1
Spot 2	1	0	1	1	1
Spot 3	1	1	0	1	1
Spot 4	1	1	1	0	1
Spot 5	1	1	1	1	0

Gene	🟢	🟠
Spot 1	4	0
Spot 2	0	4
Spot 3	0	4
...		...

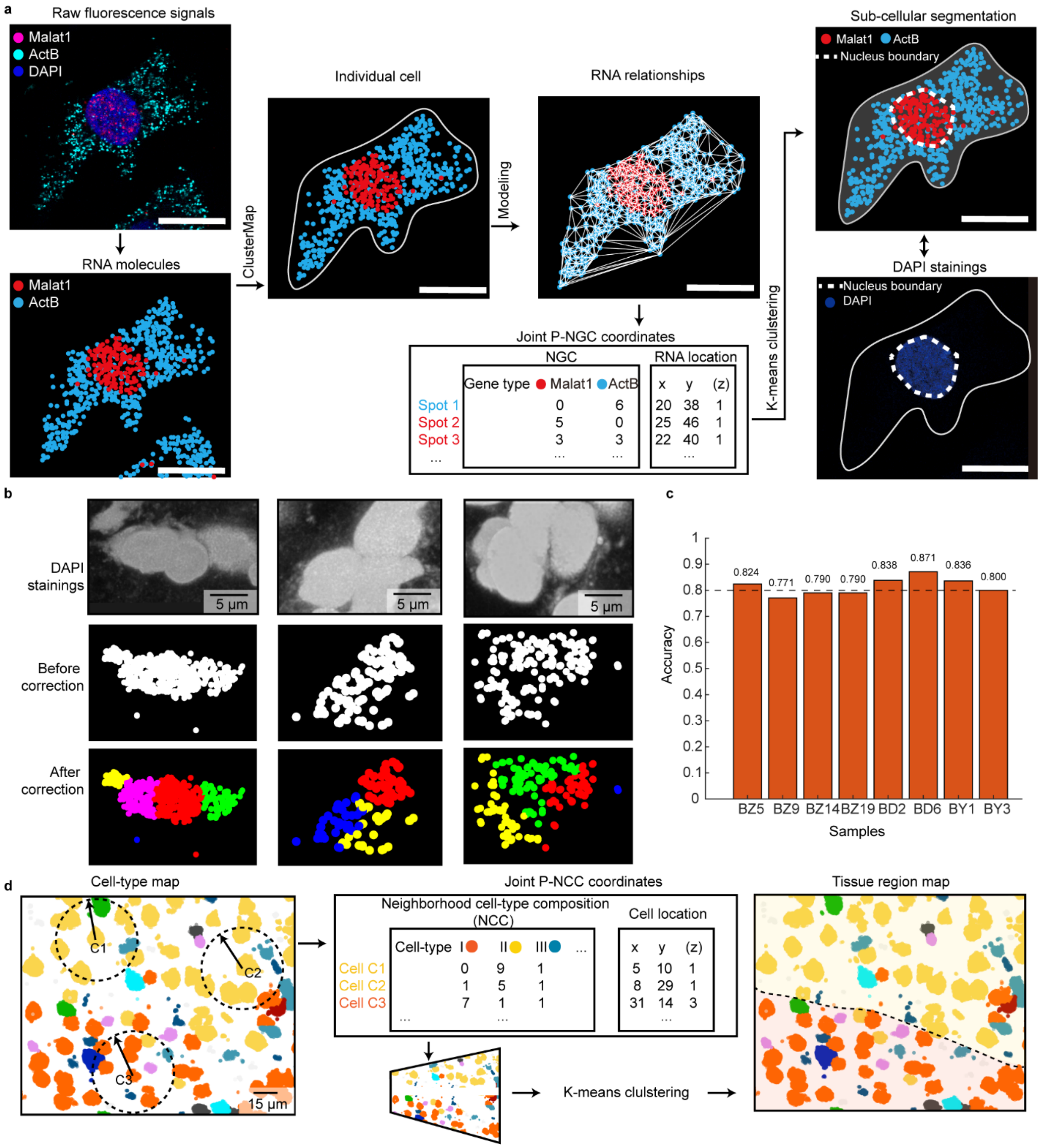
Gene	🔴	🟢	🟣	🔵	🟠
Spot 1	1	1	1	1	1
Spot 2	1	1	1	1	1
Spot 3	1	1	1	1	1
Spot 4	1	1	1	1	1
Spot 5	1	1	1	1	1

Gene	🔴	🟢	🟣	🔵	🟠
Spot 1	1	0	0	0	0
Spot 2	0	1	0	0	0
Spot 3	0	0	1	0	0
Spot 4	0	0	0	1	0
Spot 5	0	0	0	0	1

Gene	🟢	🟠
Spot 1	5	5
Spot 2	5	5
Spot 3	0	5
...		...

Supplementary Figure 2

Comparison of RNA sampling approaches between ClusterMap using absolute physical distance with other methods[20] using *k*-nearest neighbours (kNN) in simulated data. Three examples of RNAs with various local density demonstrating that ClusterMap preserves the local physical density information while kNN does not consider the physical density of RNAs.

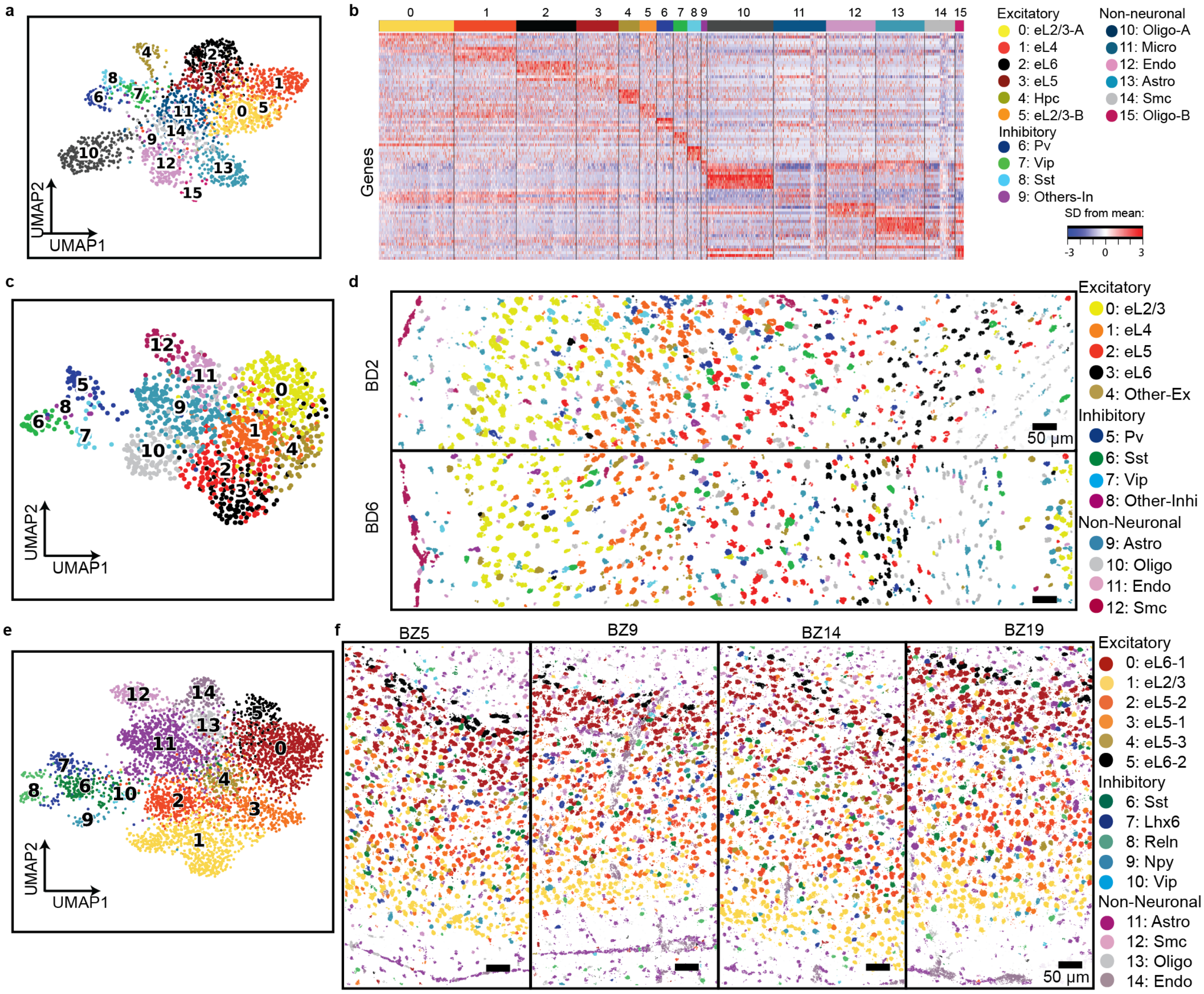


	Gene type ● Malat1	● ActB	x	y	(z)
Spot 1	0	6	20	38	1
Spot 2	5	0	25	46	1
Spot 3	3	3	22	40	1
...	...	...		...	

	Cell-type I ●	II ●	III ●	...	x	y	(z)
Cell C1	0	9	1		5	10	1
Cell C2	1	5	1		8	29	1
Cell C3	7	1	1		31	14	3
...		...				...	

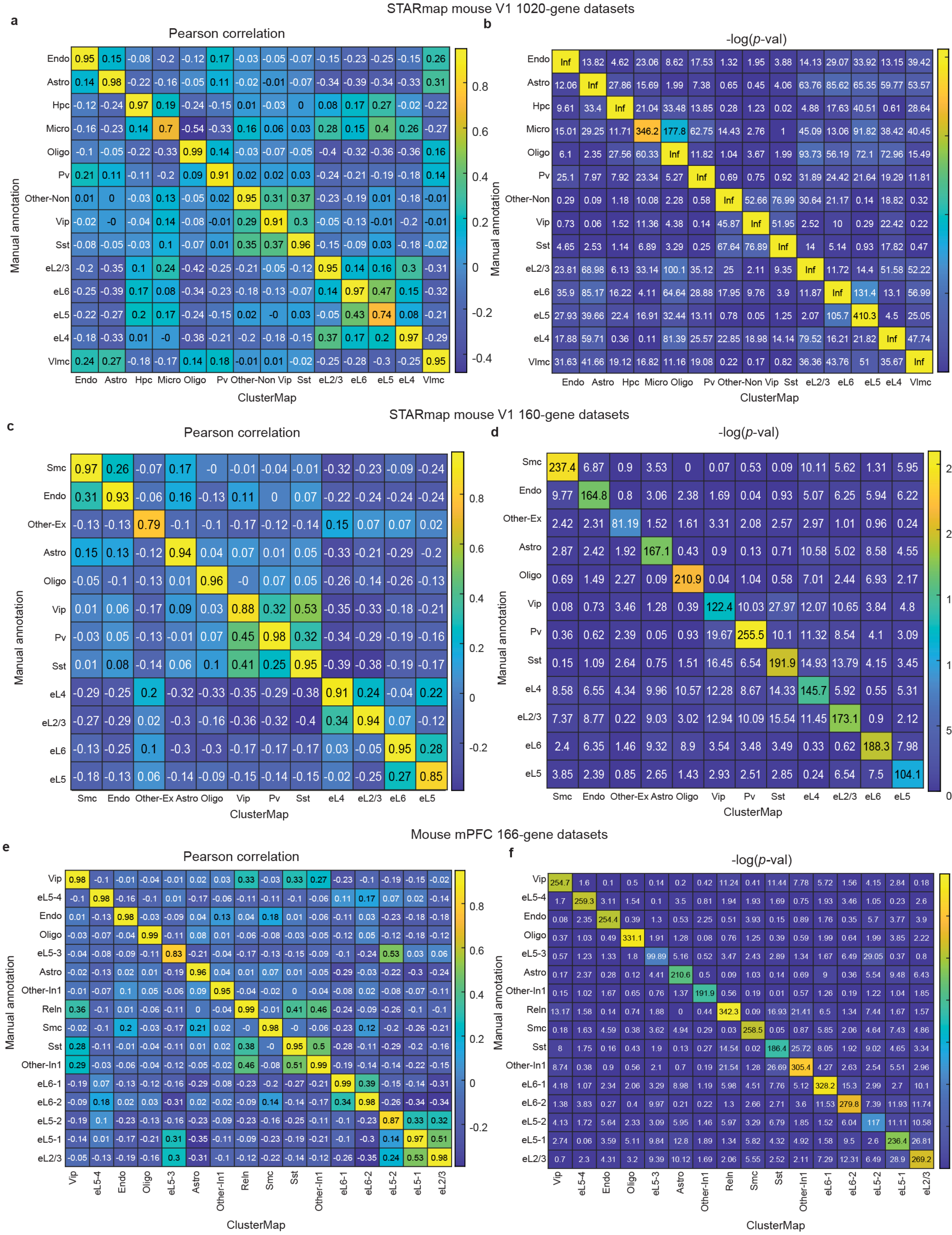
Supplementary Figure 3

Illustration of sub-cellular, cellular, and tissue region analyses. a, Subcellular analysis process of the panel IV in **Figure 1d** by ClusterMap. A three-channel (magenta: *Malat1*; cyan: *ActB*; blue: DAPI) composite image shows raw fluorescent signals. After preprocessing mRNA molecules with specific genes located, ClusterMap first performs cellular resolution and identifies individual cells. Then a mesh graph that models the relationships among mRNA spots in the cell is generated to compute the NGC coordinates and *K*-means clustering separate spots into two regions using joint physical and NGC coordinates. 100 times K-means clustering was performed with different seeds and showed the consistent same results. Finally, a convex hull is constructed from the nucleus spots, denoting the nucleus boundary. The pattern of ClusterMap-constructed nucleus boundary is compared with the DAPI staining. Scale bar: 20µm. **b**, Examples of the cell identification in ClusterMap procedures in **Fig. 2a**. Upper: DAPI staining showing the cell nuclei. Middle: mRNA spots. Lower: Clustering results. **c**, The accuracy of cell identification results from eight STARmap[6] datasets compared with corresponding expert-annotated labels. BZ5, BZ9, BZ14, BZ19: four STARmap[6] 166-gene sets in mouse medial prefrontal cortex (mPFC); BD2, BD6: two STARmap[6] 160-gene sets in mouse V1. BY1, BY3: two STARmap 1020-gene sets in mouse V1. The horizontal line is at 80% accuracy. **d**, ClusterMap constructs the tissue regions after cell-typing. First, the neighborhood cell-type composition (NCC) of each cell is computed by considering a sliding window over the cell-type map. Then both the NCC and physical locations of cells are combined for *K*-means clustering. Cells with highly correlated neighboring cell-type composition and close spatial distances are merged into a single tissue region signature.


a
UMAP2
UMAP1
b
Genes
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Excitatory
0: eL2/3-A
1: eL4
2: eL6
3: eL5
4: Hpc
5: eL2/3-B
Inhibitory
6: Pv
7: Vip
8: Sst
9: Others-In
Non-neuronal
10: Oligo-A
11: Micro
12: Endo
13: Astro
14: Smc
15: Oligo-B
SD from mean:
-3 0 3
c
UMAP2
UMAP1
d
BD2
BD6
50 µm
Excitatory
0: eL2/3
1: eL4
2: eL5
3: eL6
4: Other-Ex
Inhibitory
5: Pv
6: Sst
7: Vip
8: Other-Inhi
Non-Neuronal
9: Astro
10: Oligo
11: Endo
12: Smc
e
UMAP2
UMAP1
f
BZ5
BZ9
BZ14
BZ19
50 µm
Excitatory
0: eL6-1
1: eL2/3
2: eL5-2
3: eL5-1
4: eL5-3
5: eL6-2
Inhibitory
6: Sst
7: Lhx6
8: Reln
9: Npy
10: Vip
Non-Neuronal
11: Astro
12: Smc
13: Oligo
14: Endo

Supplementary Figure 4

Identification of cell types in mouse primary cortex (V1) and medial prefrontal cortex (mPFC). a,**b**, UMAP and heatmap visualization of all excitatory, inhibitory and non-neuronal cell types in STARmap[6] mouse V1 1020-gene (two replicates: BY1 and BY3). The heatmap represents z-scored expression matrix of all cell types, showing clustering of five most differentially expressed genes per cell type. Genes shown are selected based on a false discovery rate (FDR)-adjusted *p*-value threshold of 0.05 (Benjamini-Hochberg correction) and a minimum log_{10} fold change of 0.4, using a two-sided, unpaired t-test, for genes that are expressed in cells within each cluster versus cells in any other cluster. **c**, **e**, UMAP visualization of all excitatory, inhibitory and non-neuronal cell types in STARmap[6] 160-gene datasets in mouse V1 (two replicates: BD2, BD6, (**c**)), and STARmap[6] 166-gene datasets in mPFC (four replicates, BZ5, BZ9, BZ14, BZ19, (**e**)). **d**, **f**, Spatial organization map of cell types in BD2 and BD6 (**d**), and in BZ5, BZ9, BZ14 and BZ19 (**f**).



STARmap mouse V1 1020-gene datasets

a Pearson correlation

Manual annotation \ ClusterMap	Endo	Astro	Hpc	Micro	Oligo	Pv	Other-Non	Vip	Sst	eL2/3	eL6	eL5	eL4	Vlmc
Endo	0.95	0.15	-0.08	-0.2	-0.12	0.17	-0.03	-0.05	-0.07	-0.15	-0.23	-0.25	-0.15	0.26
Astro	0.14	0.98	-0.22	-0.16	-0.05	0.11	-0.02	-0.01	-0.07	-0.34	-0.39	-0.34	-0.33	0.31
Hpc	-0.12	-0.24	0.97	0.19	-0.24	-0.15	0.01	-0.03	0	0.08	0.17	0.27	-0.02	-0.22
Micro	-0.16	-0.23	0.14	0.7	-0.54	-0.33	0.16	0.06	0.03	0.28	0.15	0.4	0.26	-0.27
Oligo	-0.1	-0.05	-0.22	-0.33	0.99	0.14	-0.03	-0.07	-0.05	-0.4	-0.32	-0.36	-0.36	0.16
Pv	0.21	0.11	-0.11	-0.2	0.09	0.91	0.02	0.02	0.03	-0.24	-0.21	-0.19	-0.18	0.14
Other-Non	0.01	0	-0.03	0.13	-0.05	0.02	0.95	0.31	0.37	-0.23	-0.19	0.01	-0.18	-0.01
Vip	-0.02	-0	-0.04	0.14	-0.08	-0.01	0.29	0.91	0.3	-0.05	-0.13	-0.01	-0.2	-0.01
Sst	-0.08	-0.05	-0.03	0.1	-0.07	0.01	0.35	0.37	0.96	-0.15	-0.09	0.03	-0.18	-0.02
eL2/3	-0.2	-0.35	0.1	0.24	-0.42	-0.25	-0.21	-0.05	-0.12	0.95	0.14	0.16	0.3	-0.31
eL6	-0.25	-0.39	0.17	0.08	-0.34	-0.23	-0.18	-0.13	-0.07	0.14	0.97	0.47	0.15	-0.32
eL5	-0.22	-0.27	0.2	0.17	-0.24	-0.15	0.02	-0	0.03	-0.05	0.43	0.74	0.08	-0.21
eL4	-0.18	-0.33	0.01	-0	-0.38	-0.21	-0.2	-0.18	-0.15	0.37	0.17	0.2	0.97	-0.29
Vlmc	0.24	0.27	-0.18	-0.17	0.14	0.18	-0.01	0.01	-0.02	-0.25	-0.28	-0.3	-0.25	0.95

b -log(*p*-val)

Manual annotation \ ClusterMap	Endo	Astro	Hpc	Micro	Oligo	Pv	Other-Non	Vip	Sst	eL2/3	eL6	eL5	eL4	Vlmc
Endo	Inf	13.82	4.62	23.06	8.62	17.53	1.32	1.95	3.88	14.13	29.07	33.92	13.15	39.42
Astro	12.06	Inf	27.86	15.69	1.99	7.38	0.65	0.45	4.06	63.76	85.62	65.35	59.77	53.57
Hpc	9.61	33.4	Inf	21.04	33.48	13.85	0.28	1.23	0.02	4.88	17.63	40.51	0.61	28.64
Micro	15.01	29.25	11.71	346.2	177.8	62.75	14.43	2.76	1	45.09	13.06	91.82	38.42	40.45
Oligo	6.1	2.35	27.56	60.33	Inf	11.82	1.04	3.67	1.99	93.73	56.19	72.1	72.96	15.49
Pv	25.1	7.97	7.92	23.34	5.27	Inf	0.69	0.75	0.92	31.89	24.42	21.64	19.29	11.81
Other-Non	0.29	0.09	1.18	10.08	2.28	0.58	Inf	52.66	76.99	30.64	21.17	0.14	18.82	0.32
Vip	0.73	0.06	1.52	11.36	4.38	0.14	45.87	Inf	51.95	2.52	10	0.29	22.42	0.22
Sst	4.65	2.53	1.14	6.89	3.29	0.25	67.64	76.89	Inf	14	5.14	0.93	17.82	0.47
eL2/3	23.81	68.98	6.13	33.14	100.1	35.12	25	2.11	9.35	Inf	11.72	14.4	51.58	52.22
eL6	35.9	85.17	16.22	4.11	64.64	28.88	17.95	9.76	3.9	11.87	Inf	131.4	13.1	56.99
eL5	27.93	39.66	22.4	16.91	32.44	13.11	0.78	0.05	1.25	2.07	105.7	410.3	4.5	25.05
eL4	17.88	59.71	0.36	0.11	81.39	25.57	22.85	18.98	14.14	79.52	16.21	21.82	Inf	47.74
Vlmc	31.63	41.66	19.12	16.82	11.16	19.08	0.22	0.17	0.82	36.36	43.76	51	35.67	Inf

STARmap mouse V1 160-gene datasets

c Pearson correlation

Manual annotation \ ClusterMap	Smc	Endo	Other-Ex	Astro	Oligo	Vip	Pv	Sst	eL4	eL2/3	eL6	eL5
Smc	0.97	0.26	-0.07	0.17	-0	-0.01	-0.04	-0.01	-0.32	-0.23	-0.09	-0.24
Endo	0.31	0.93	-0.06	0.16	-0.13	0.11	0	0.07	-0.22	-0.24	-0.24	-0.24
Other-Ex	-0.13	-0.13	0.79	-0.1	-0.1	-0.17	-0.12	-0.14	0.15	0.07	0.07	0.02
Astro	0.15	0.13	-0.12	0.94	0.04	0.07	0.01	0.05	-0.33	-0.21	-0.29	-0.2
Oligo	-0.05	-0.1	-0.13	0.01	0.96	-0	0.07	0.05	-0.26	-0.14	-0.26	-0.13
Vip	0.01	0.06	-0.17	0.09	0.03	0.88	0.32	0.53	-0.35	-0.33	-0.18	-0.21
Pv	-0.03	0.05	-0.13	-0.01	0.07	0.45	0.98	0.32	-0.34	-0.29	-0.19	-0.16
Sst	0.01	0.08	-0.14	0.06	0.1	0.41	0.25	0.95	-0.39	-0.38	-0.19	-0.17
eL4	-0.29	-0.25	0.2	-0.32	-0.33	-0.35	-0.29	-0.38	0.91	0.24	-0.04	0.22
eL2/3	-0.27	-0.29	0.02	-0.3	-0.16	-0.36	-0.32	-0.4	0.34	0.94	0.07	-0.12
eL6	-0.13	-0.25	0.1	-0.3	-0.3	-0.17	-0.17	-0.17	0.03	-0.05	0.95	0.28
eL5	-0.18	-0.13	0.06	-0.14	-0.09	-0.15	-0.14	-0.15	-0.02	-0.25	0.27	0.85

d -log(*p*-val)

Manual annotation \ ClusterMap	Smc	Endo	Other-Ex	Astro	Oligo	Vip	Pv	Sst	eL4	eL2/3	eL6	eL5
Smc	237.4	6.87	0.9	3.53	0	0.07	0.53	0.09	10.11	5.62	1.31	5.95
Endo	9.77	164.8	0.8	3.06	2.38	1.69	0.04	0.93	5.07	6.25	5.94	6.22
Other-Ex	2.42	2.31	81.19	1.52	1.61	3.31	2.08	2.57	2.97	1.01	0.96	0.24
Astro	2.87	2.42	1.92	167.1	0.43	0.9	0.13	0.71	10.58	5.02	8.58	4.55
Oligo	0.69	1.49	2.27	0.09	210.9	0.04	1.04	0.58	7.01	2.44	6.93	2.17
Vip	0.08	0.73	3.46	1.28	0.39	122.4	10.03	27.97	12.07	10.65	3.84	4.8
Pv	0.36	0.62	2.39	0.05	0.93	19.67	255.5	10.1	11.32	8.54	4.1	3.09
Sst	0.15	1.09	2.64	0.75	1.51	16.45	6.54	191.9	14.93	13.79	4.15	3.45
eL4	8.58	6.55	4.34	9.96	10.57	12.28	8.67	14.33	145.7	5.92	0.55	5.31
eL2/3	7.37	8.77	0.22	9.03	3.02	12.94	10.09	15.54	11.45	173.1	0.9	2.12
eL6	2.4	6.35	1.46	9.32	8.9	3.54	3.48	3.49	0.33	0.62	188.3	7.98
eL5	3.85	2.39	0.85	2.65	1.43	2.93	2.51	2.85	0.24	6.54	7.5	104.1

Mouse mPFC 166-gene datasets

e Pearson correlation

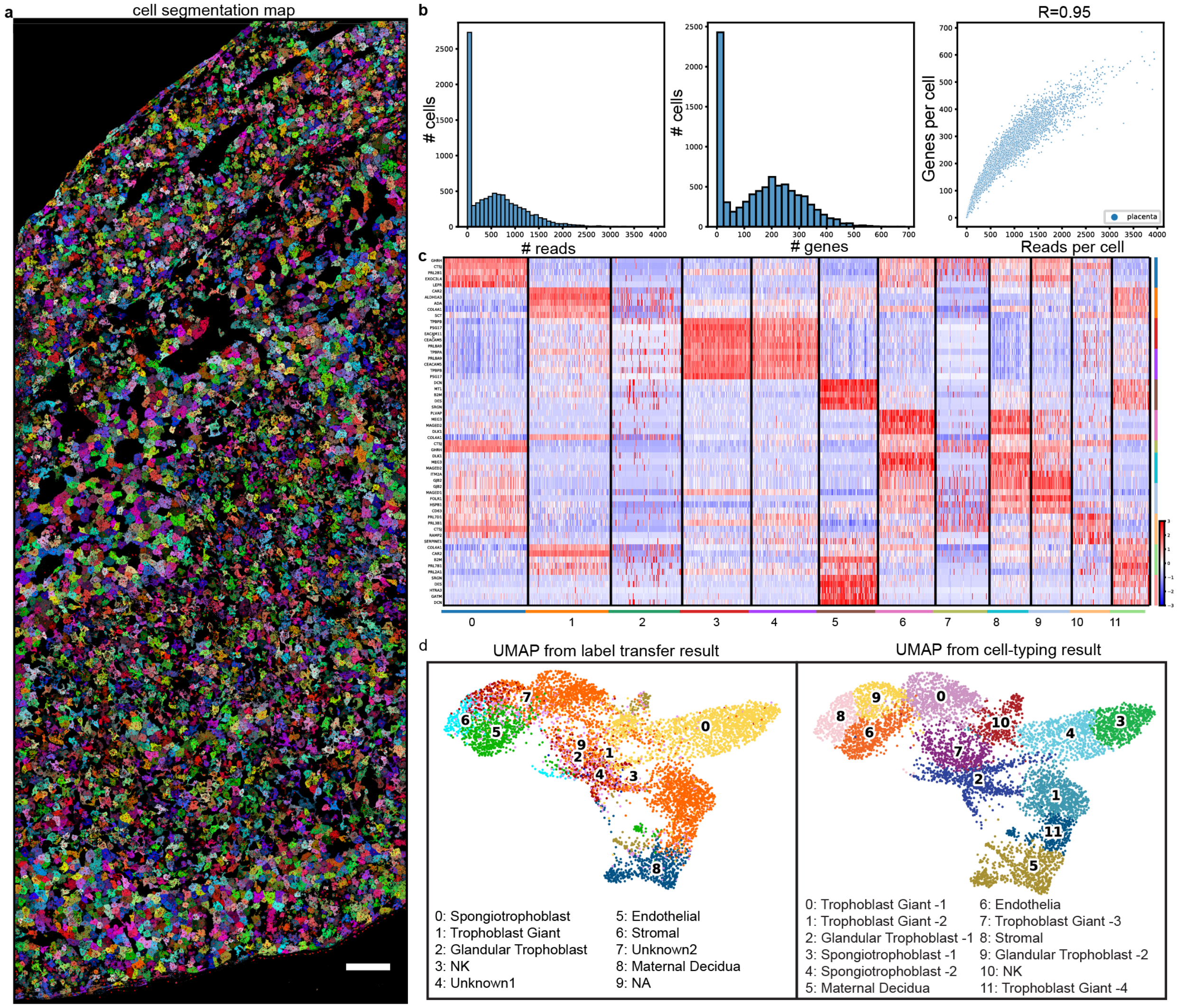
Manual annotation \ ClusterMap	Vip	eL5-4	Endo	Oligo	eL5-3	Astro	Other-In1	Reln	Smc	Sst	Other-In1	eL6-1	eL6-2	eL5-2	eL5-1	eL2/3
Vip	0.98	-0.1	-0.01	-0.04	-0.01	0.02	0.03	0.33	-0.03	0.33	0.27	-0.23	-0.1	-0.19	-0.15	-0.02
eL5-4	-0.1	0.98	-0.16	-0.1	0.01	-0.17	-0.06	-0.11	-0.11	-0.1	-0.06	0.11	0.17	0.07	0.02	-0.14
Endo	0.01	-0.13	0.98	-0.03	-0.09	0.04	0.13	0.04	0.18	0.01	-0.06	-0.11	0.03	-0.23	-0.18	-0.18
Oligo	-0.03	-0.07	-0.04	0.99	-0.11	0.08	0.01	-0.06	-0.08	-0.03	-0.05	-0.12	0.05	-0.12	-0.18	-0.12
eL5-3	-0.04	-0.08	-0.09	-0.11	0.83	-0.21	-0.04	-0.17	-0.13	-0.15	-0.09	-0.1	-0.24	0.53	0.03	0.06
Astro	-0.02	-0.13	0.02	0.01	-0.19	0.96	0.04	0.01	0.07	0.01	-0.05	-0.29	-0.03	-0.22	-0.3	-0.24
Other-In1	-0.01	-0.07	0.1	0.05	-0.06	0.09	0.95	-0.04	-0.02	0	-0.04	-0.08	0.02	-0.08	-0.07	-0.11
Reln	0.36	-0.1	0.01	-0.06	-0.11	0	-0.04	0.99	-0.01	0.41	0.46	-0.24	-0.09	-0.26	-0.1	-0.1
Smc	-0.02	-0.1	0.2	-0.03	-0.17	0.21	0.02	-0	0.98	-0	-0.06	-0.23	0.12	-0.2	-0.26	-0.21
Sst	0.28	-0.11	-0.01	-0.04	-0.11	0.01	0.02	0.38	-0	0.95	0.5	-0.28	-0.11	-0.29	-0.2	-0.16
Other-In1	0.29	-0.03	-0.06	-0.04	-0.12	-0.05	-0.02	0.46	-0.08	0.51	0.99	-0.19	-0.14	-0.14	-0.22	-0.15
eL6-1	-0.19	0.07	-0.13	-0.12	-0.16	-0.29	-0.08	-0.23	-0.2	-0.27	-0.21	0.99	0.39	-0.15	-0.14	-0.31
eL6-2	-0.09	0.18	0.02	0.03	-0.31	0.02	-0.02	-0.09	0.14	-0.14	-0.17	0.34	0.98	-0.26	-0.34	-0.34
eL5-2	-0.19	0.1	-0.23	-0.13	-0.16	-0.23	-0.09	-0.23	-0.16	-0.25	-0.11	-0.1	-0.23	0.87	0.33	0.32
eL5-1	-0.14	0.01	-0.17	-0.21	0.31	-0.35	-0.11	-0.09	-0.23	-0.19	-0.21	-0.1	-0.3	0.14	0.97	0.51
eL2/3	-0.05	-0.13	-0.19	-0.16	0.3	-0.31	-0.1	-0.12	-0.22	-0.14	-0.12	-0.26	-0.35	0.24	0.53	0.98

f -log(*p*-val)

Manual annotation \ ClusterMap	Vip	eL5-4	Endo	Oligo	eL5-3	Astro	Other-In1	Reln	Smc	Sst	Other-In1	eL6-1	eL6-2	eL5-2	eL5-1	eL2/3
Vip	254.7	1.6	0.1	0.5	0.14	0.2	0.42	11.24	0.41	11.44	7.78	5.72	1.56	4.15	2.84	0.18
eL5-4	1.7	259.3	3.11	1.54	0.1	3.5	0.81	1.94	1.93	1.69	0.75	1.93	3.46	1.05	0.23	2.6
Endo	0.08	2.35	254.4	0.39	1.3	0.53	2.25	0.51	3.93	0.15	0.89	1.76	0.35	5.7	3.77	3.9
Oligo	0.37	1.03	0.49	331.1	1.91	1.28	0.08	0.76	1.25	0.39	0.59	1.99	0.64	1.99	3.85	2.22
eL5-3	0.57	1.23	1.33	1.8	99.89	5.16	0.52	3.47	2.43	2.89	1.34	1.67	6.49	29.05	0.37	0.8
Astro	0.17	2.37	0.28	0.12	4.41	210.6	0.5	0.09	1.03	0.14	0.69	9	0.36	5.54	9.48	6.43
Other-In1	0.15	1.02	1.67	0.65	0.76	1.37	191.9	0.56	0.19	0.01	0.57	1.26	0.19	1.22	1.04	1.85
Reln	13.17	1.58	0.14	0.74	1.88	0	0.44	342.3	0.09	16.93	21.41	6.5	1.34	7.44	1.67	1.57
Smc	0.18	1.63	4.59	0.38	3.62	4.94	0.29	0.03	258.5	0.05	0.87	5.85	2.06	4.64	7.43	4.86
Sst	8	1.75	0.16	0.43	1.9	0.13	0.27	14.54	0.02	186.4	25.72	8.05	1.92	9.02	4.65	3.34
Other-In1	8.74	0.38	0.9	0.56	2.1	0.7	0.19	21.54	1.28	26.69	305.4	4.27	2.63	2.54	5.51	2.96
eL6-1	4.18	1.07	2.34	2.06	3.29	8.98	1.19	5.98	4.51	7.76	5.12	328.2	15.3	2.99	2.7	10.1
eL6-2	1.38	3.83	0.27	0.4	9.97	0.21	0.22	1.3	2.66	2.71	3.6	11.53	279.8	7.39	11.93	11.74
eL5-2	4.13	1.72	5.64	2.33	3.09	5.95	1.46	5.97	3.29	6.79	1.85	1.52	6.04	117	11.11	10.58
eL5-1	2.74	0.06	3.59	5.11	9.84	12.8	1.89	1.34	5.82	4.32	4.92	1.58	9.5	2.6	236.4	26.81
eL2/3	0.7	2.3	4.31	3.2	9.39	10.12	1.69	2.06	5.55	2.52	2.11	7.29	12.31	6.49	28.9	269.2

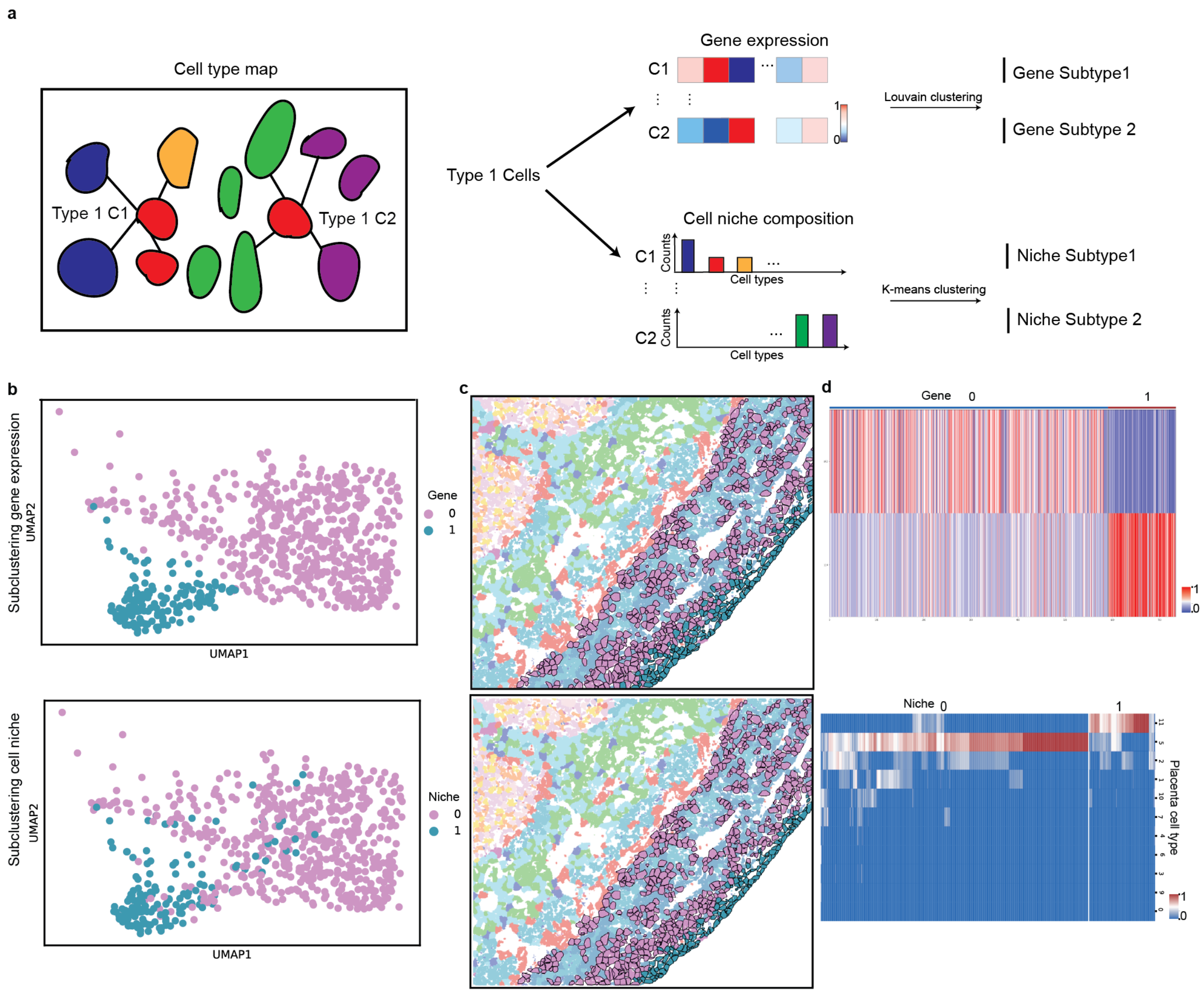
Supplementary Figure 5

Cell-type correlation matrices comparison of ClusterMap-based and manually-segmented cell types. **a**,**b**, Comparison on STARmap mouse V1 1020-gene datasets. Heatmaps of Pearson correlation (**a**) and *-log*(*p*-value) (**b**) for null hypothesis testing. The *p* value is based on a t statistic which has *n*-2 degrees of freedom and 95% confidence interval. The single-cell gene expression profiles from ClusterMap with manual annotation are compared. **c**,**d**, Comparison on STARmap mouse V1 160-gene datasets. Heatmaps of correlation (**c**) and *-log*(*p*-value) (**d**) comparing the single-cell gene expression profiles from ClusterMap with manual annotation. **e**,**f**, Comparison on STARmap mouse mPFC 166-gene datasets. Heatmaps of correlation (**e**) and *-log*(*p*-value) (**f**) comparing the single-cell gene expression profiles from ClusterMap with manual annotation. Horizontal: ClusterMap; vertical: manual annotation.


a
cell segmentation map
b
# cells
# reads
0
500
1000
1500
2000
2500
3000
3500
4000
# cells
# genes
0
100
200
300
400
500
600
700
R=0.95
Genes per cell
Reads per cell
placenta
c
GHRH
CTSJ
PRL2B1
EXOC3L4
LEPR
CAR2
ALDH1A3
ADA
COL4A1
SCT
TPBPB
PSG17
EACAM11
CEACAM5
PRL8A9
TPBPA
PRL8A9
CEACAM5
TPBPB
PSG17
DCN
MT1
B2M
DES
SRGN
PLVAP
MEG3
MAGED2
DLK1
COL4A1
CTSJ
GHRH
DLK1
MEG3
MAGED2
ITM2A
GJB2
GJB2
MAGED1
FOLR1
HSPB1
CD63
PRL7D1
PRL3B1
CTSJ
RAMP2
SERPINE1
COL4A1
CAR2
B2M
PRL7B1
PRL2A1
SRGN
DES
HTRA3
GATM
DCN
0
1
2
3
4
5
6
7
8
9
10
11
d
UMAP from label transfer result
0: Spongiotrophoblast
1: Trophoblast Giant
2: Glandular Trophoblast
3: NK
4: Unknown1
5: Endothelial
6: Stromal
7: Unknown2
8: Maternal Decidua
9: NA
UMAP from cell-typing result
0: Trophoblast Giant -1
1: Trophoblast Giant -2
2: Glandular Trophoblast -1
3: Spongiotrophoblast -1
4: Spongiotrophoblast -2
5: Maternal Decidua
6: Endothelia
7: Trophoblast Giant -3
8: Stromal
9: Glandular Trophoblast -2
10: NK
11: Trophoblast Giant -4

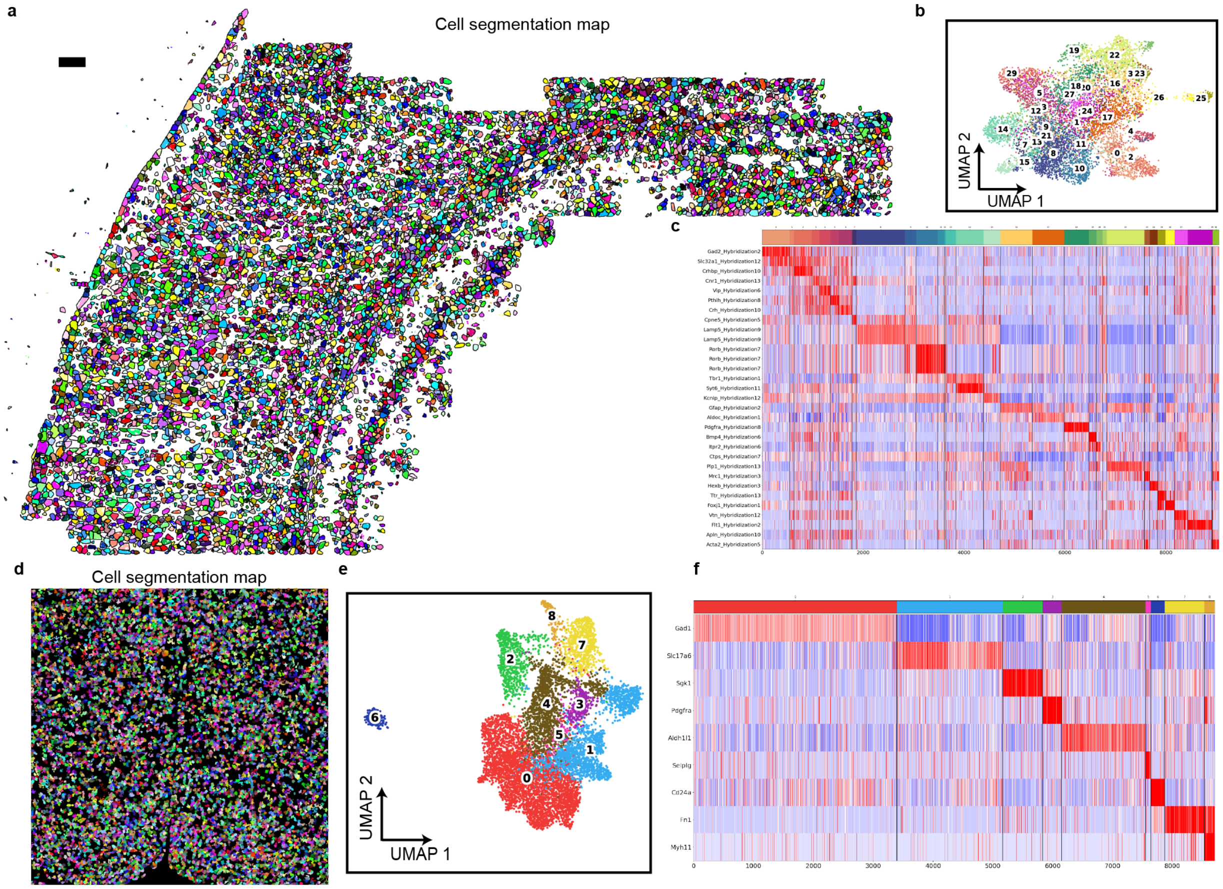
Supplementary Figure 6

Analyses of the STARmap[6] mouse placental dataset. **a**, ClusterMap generates the cell segmentation map of the STARmap[6] mouse placenta 903-gene dataset, including 7,224 cells. Scale bar: 100 µm. **b**, Statistics of ClusterMap identified placental cells as shown in (**a**). Left: Histogram of detected reads (DNA amplicons) per cell. Middle: Histogram of genes per cell. Right: Correlation plot between genes per cell and reads per cell. **c**, Heatmap visualization of 12 cell types. Names are in the right panel of (**d**). **d**, UMAP from label transfer results with scRNA-seq, compared with UMAP of the Louvain clustering[22] in ClusterMap.


a
Cell type map
Type 1 C1
Type 1 C2
Type 1 Cells
Gene expression
C1
C2
1
0
Louvain clustering
Gene Subtype1
Gene Subtype 2
Cell niche composition
Counts
Cell types
K-means clustering
Niche Subtype1
Niche Subtype 2
b
Subclustering gene expression
UMAP2
UMAP1
Gene
0
1
Subclustering cell niche
UMAP2
UMAP1
Niche
0
1
c
d
Gene 0
1
1
0
Niche 0
1
Placenta cell type
1
0

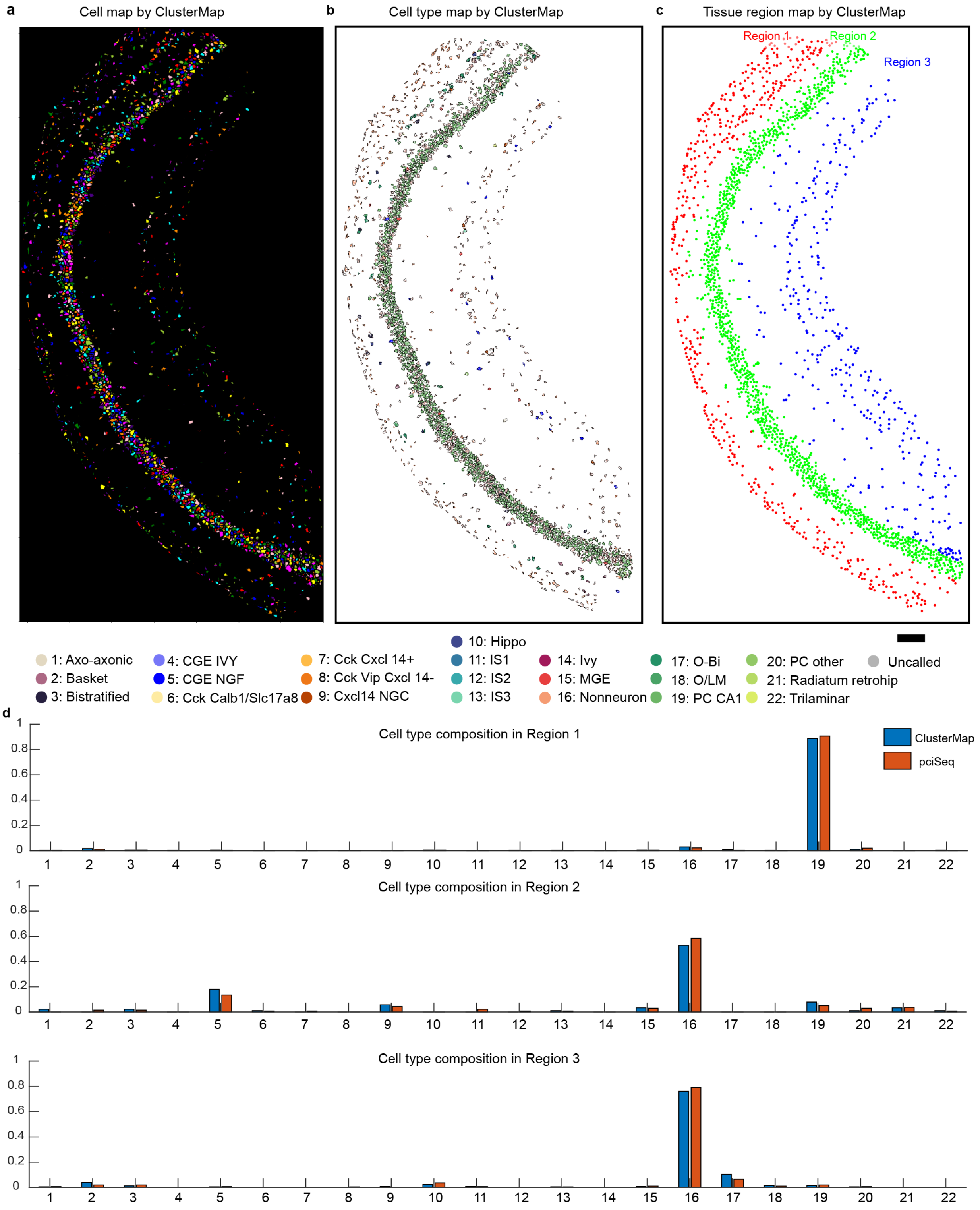
Supplementary Figure 7

Sub-clustering within one cell type using cell niche compositions in STARmap mouse placenta 903-gene dataset. a, Schematic indicating how cells in one cell type are sub-clustered based on either gene expression (Louvain clustering[22]) or the cell niche compositions (*K*-means clustering[19]). **b**, UMAP of gene expression sub-clustering (top) or cell niche composition sub-clustering (bottom) in Maternal Decidua-1 (MD-1). **c**, Spatial subtype maps using gene expression (top) or cell niche composition (bottom) in MD-1. **d**, Heatmap of sub-clustering using gene expression (top) or cell niche composition sub-clustering (bottom) in MD-1. Gene markers in the top heatmaps of gene expression sub-clustering are 0: *GPNMB*, 1: *CXCL14*. Row names in the bottom heatmaps of cell niche composition sub-clustering are cell types in numbers annotated in Figure 3.


a
Cell segmentation map
b
UMAP 2
UMAP 1
19
22
29
3 23
18
10
16
26
25
5
27
12
3
24
17
14
9
21
4
7
13
8
11
0
2
15
10
c
Gad2_Hybridization2
Slc32a1_Hybridization12
Crhbp_Hybridization10
Cnr1_Hybridization13
Vip_Hybridization6
Pthlh_Hybridization8
Crh_Hybridization10
Cpne5_Hybridization5
Lamp5_Hybridization9
Lamp5_Hybridization9
Rorb_Hybridization7
Rorb_Hybridization7
Rorb_Hybridization7
Tbr1_Hybridization1
Syt6_Hybridization11
Kcnip_Hybridization12
Gfap_Hybridization2
Aldoc_Hybridization1
Pdgfra_Hybridization8
Bmp4_Hybridization6
Itpr2_Hybridization6
Ctps_Hybridization7
Plp1_Hybridization13
Mrc1_Hybridization3
Hexb_Hybridization3
Ttr_Hybridization13
Foxj1_Hybridization1
Vtn_Hybridization12
Flt1_Hybridization2
Apln_Hybridization10
Acta2_Hybridization5
0
2000
4000
6000
8000
d
Cell segmentation map
e
8
7
2
4
3
6
5
1
0
UMAP 2
UMAP 1
f
Gad1
Slc17a6
Sgk1
Pdgfra
Aldh1l1
Selplg
Cd24a
Fn1
Myh11
0
1000
2000
3000
4000
5000
6000
7000
8000

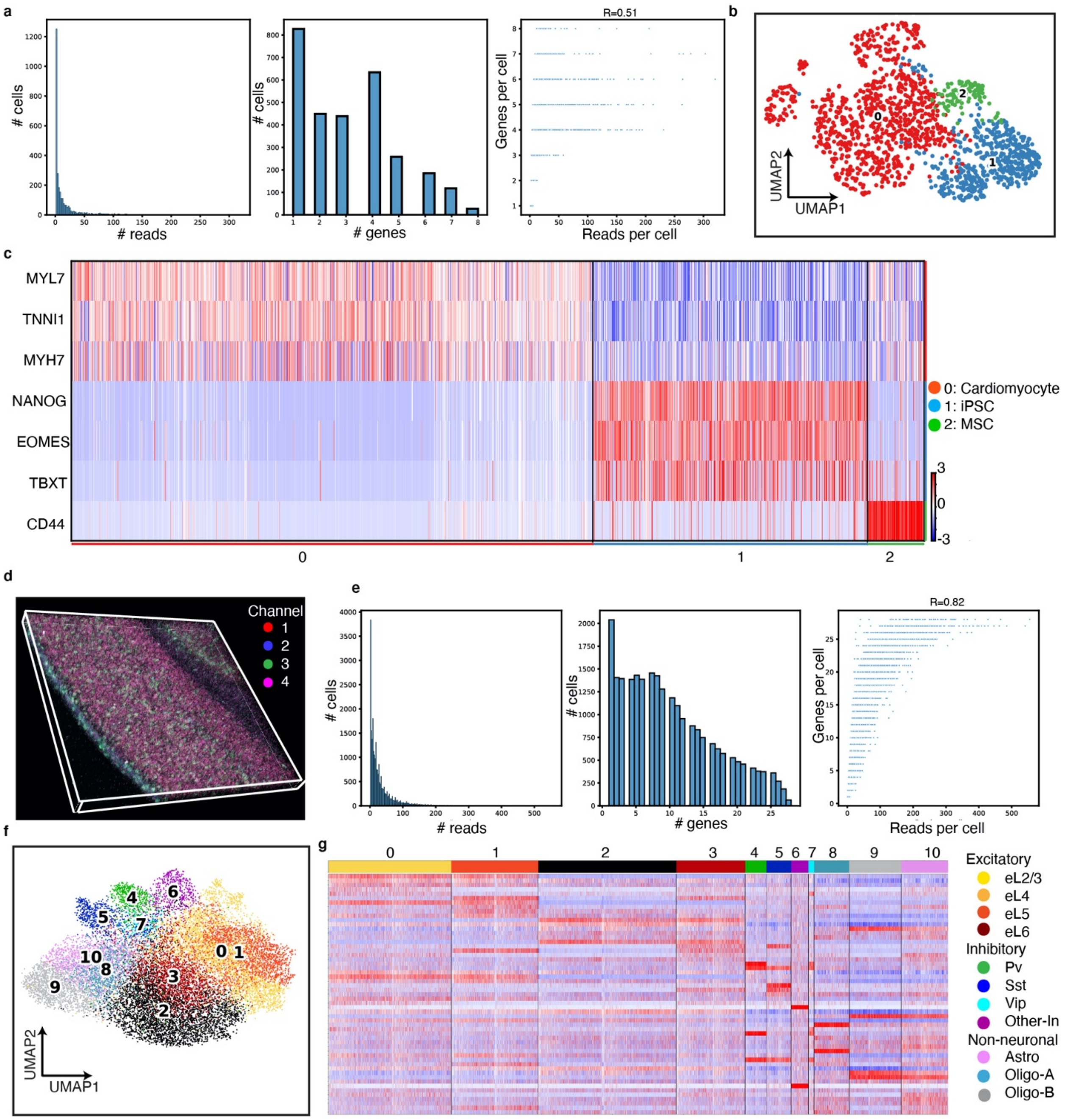
Supplementary Figure 8

ClusterMap analyses across different experimental methods. a, The cell segmentation map of whole osmFISH mouse somatosensory cortex (SSp) datasets. Scale bar: 100 µm. **b,c**, UMAP and heatmap visualization of 31 cell types in osmFISH datasets. **d**, The 2D cell segmentation map of whole MERFISH mouse preoptic area (POA) datasets. Scale bar: 200 µm. **e**,**f**, UMAP and heatmap visualization of 9 cell types in MERFISH datasets. The number of cells increased from 6,471 to 8,538 for osmFISH, from 2,620 to 2,924 for pciSeq, from 6,977 to 10,320 for MERFISH. The number of reads increased from 1,248,106 to 1,690,328 for osmFISH, from 31,246 to 31,750 for pciSeq, from 1,927,913 to 3,065,171 for MERFISH.


a
Cell map by ClusterMap
b
Cell type map by ClusterMap
c
Tissue region map by ClusterMap
Region 1
Region 2
Region 3
1: Axo-axonic
2: Basket
3: Bistratified
4: CGE IVY
5: CGE NGF
6: Cck Calb1/Slc17a8
7: Cck Cxcl 14+
8: Cck Vip Cxcl 14-
9: Cxcl14 NGC
10: Hippo
11: IS1
12: IS2
13: IS3
14: Ivy
15: MGE
16: Nonneuron
17: O-Bi
18: O/LM
19: PC CA1
20: PC other
21: Radiatum retrohip
22: Trilaminar
Uncalled
d
Cell type composition in Region 1
Cell type composition in Region 2
Cell type composition in Region 3
ClusterMap
pciSeq
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

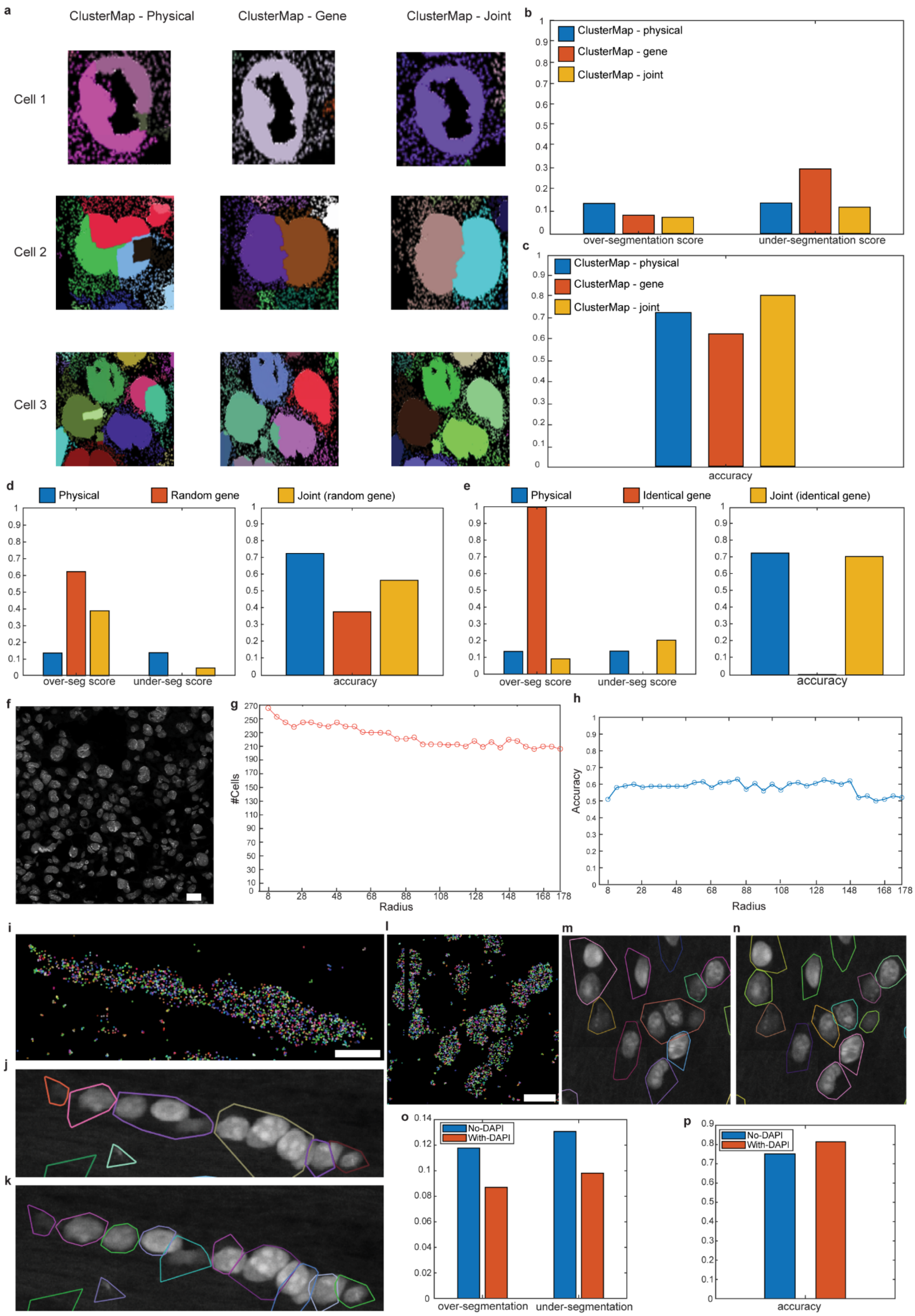
Supplementary Figure 9

ClusterMap analyses of ISS data. a, Cell segmentation map shows the cell segmentation results by ClusterMap. Colors are randomly assigned to each cell mask. **b**, Cell type map shows the cell type calling results. Colors are assigned according to their corresponding cell type categories. **c**, Tissue region map shows laminar structure of hippocampus. Scale bar: 200 μm. **d**, Side-by-side comparison of cell type compositions in each tissue region from ClusterMap and pciSeq of the ISS data.


a
# cells
# reads
# cells
# genes
R=0.51
Genes per cell
Reads per cell
b
0
1
2
UMAP2
UMAP1
c
MYL7
TNNI1
MYH7
NANOG
EOMES
TBXT
CD44
0: Cardiomyocyte
1: iPSC
2: MSC
3
0
-3
0
1
2
d
Channel
1
2
3
4
e
# cells
# reads
# cells
# genes
R=0.82
Genes per cell
Reads per cell
f
0
1
2
3
4
5
6
7
8
9
10
UMAP2
UMAP1
g
0
1
2
3
4
5
6
7
8
9
10
Excitatory
eL2/3
eL4
eL5
eL6
Inhibitory
Pv
Sst
Vip
Other-In
Non-neuronal
Astro
Oligo-A
Oligo-B

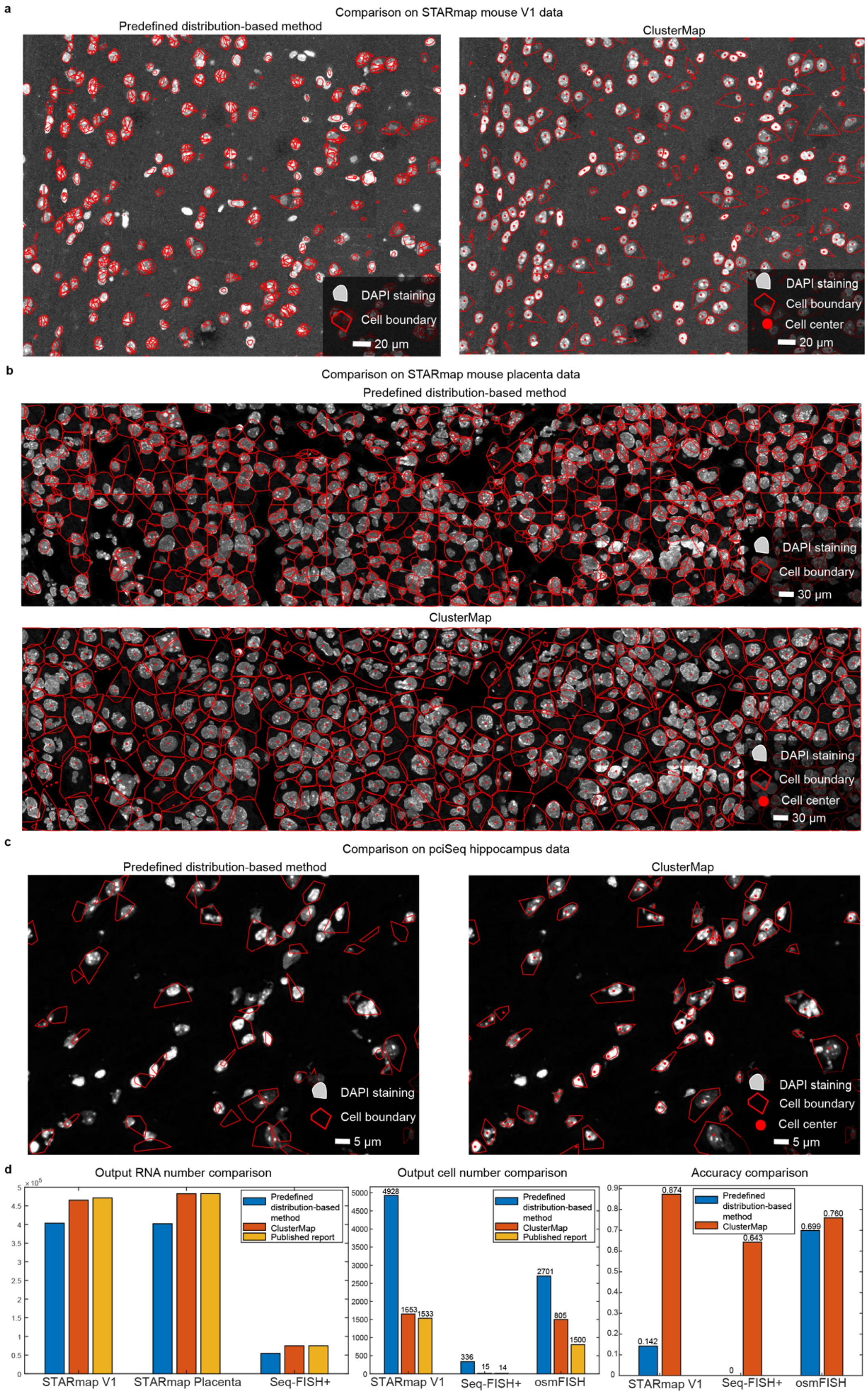
Supplementary Figure 10

ClusterMap analyses in the 3D datasets. **a**, Statistics of ClusterMap identified cells in the 3D STARmap[6] cardiac organoid[27] 8-gene dataset. Left: Histogram of detected reads (DNA amplicons) per cell. Middle: Histogram of genes per cell. Right: Correlation plot between genes per cell and reads per cell. **b**, **c**, UMAP and heatmap visualization of three cell types in the STARmap[6] cardiac organoid 8-gene dataset. The number of cells in each cell type is as follows: cardiomyocytes, 929; induced pluripotent stem cells (iPSCs), 489; mesenchymal stem cells (MSCs), 101. **d**, 3D four-channel composite raw fluorescent image of the first sequencing round shows spatial arrangement of mRNA molecules in the STARmap[6] mouse V1 28-gene dataset. Width 184 µm, height 194 µm, depth 100 µm. **e**, Statistics of ClusterMap identified cells in (**d**). Left: Histogram of detected reads (DNA amplicons) per cell. Middle: Histogram of genes per cell. Right: Correlation plot between genes per cell and reads per cell. **f**, **g**, UMAP and heatmap visualization of three cell types of (**d**).


a
ClusterMap - Physical
ClusterMap - Gene
ClusterMap - Joint
Cell 1
Cell 2
Cell 3
b
ClusterMap - physical
ClusterMap - gene
ClusterMap - joint
over-segmentation score
under-segmentation score
c
ClusterMap - physical
ClusterMap - gene
ClusterMap - joint
accuracy
d
Physical
Random gene
Joint (random gene)
over-seg score
under-seg score
accuracy
e
Physical
Identical gene
Joint (identical gene)
over-seg score
under-seg score
accuracy
f
g
#Cells
Radius
h
Accuracy
Radius
i
j
k
l
m
n
o
No-DAPI
With-DAPI
over-segmentation
under-segmentation
p
No-DAPI
With-DAPI
accuracy

Supplementary Figure 11

Performance comparison of ClusterMap using physical density, gene distance, and joint information. a, Examples of cell segmentation using only physical density information (left), gene (NGC) distance information (middle), and joint information (right). **b**,**c**, Bar plots demonstrating the percentage of over-/under- segmented cells in ground truth cells (**b**) and overall accuracy (**c**) in using physical distances information, gene (NGC) distance information, and joint information. **d**,**e**, Bar plots demonstrating the percentage of over-/under- segmented cells in ground truth cells (left) and overall accuracy (right) in using physical distances information, random (**d**) or identical (**e**) gene (NGC) distance information, and joint information. **f**, Raw DAPI image of the targeted mouse placenta tissue. Scale bar: 20μm. **g**,**h**, Line plots showing the number of cells and overall accuracy to the radius. **i-m**, Two examples of the hippocampus regions in STARmap mouse V1 1020-gene datasets showing raw spatial transcriptomics data (**i**,**l**), ClusterMap results without DAPI (**j**,**m**), and ClusterMap results with DAPI (**k**,**n**). Scale bar: 20 μm. **o**,**p**, Bar plots showing the percentage of over-/under- segmented cells (**o**) and overall accuracy (**p**) from ClusterMap without and with DAPI.


a
Comparison on STARmap mouse V1 data
Predefined distribution-based method
ClusterMap
DAPI staining
Cell boundary
Cell center
20 µm
b
Comparison on STARmap mouse placenta data
Predefined distribution-based method
ClusterMap
DAPI staining
Cell boundary
Cell center
30 µm
c
Comparison on pciSeq hippocampus data
Predefined distribution-based method
ClusterMap
DAPI staining
Cell boundary
Cell center
5 µm
d
Output RNA number comparison
×10^5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Predefined distribution-based method
ClusterMap
Published report
STARmap V1
STARmap Placenta
Seq-FISH+
Output cell number comparison
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
4928
1653
1533
336
15
14
2701
805
1500
STARmap V1
Seq-FISH+
osmFISH
Accuracy comparison
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.142
0.874
0
0.643
0.699
0.760
STARmap V1
Seq-FISH+
osmFISH

Supplementary Figure 12

Performance comparison of ClusterMap and other method across different types of *in situ* transcriptomic data. a, Example of a region in the STARmap[6] mouse V1 1020-gene dataset with DAPI signals (gray) showing ground truth cell nuclei. Red contours show cell boundaries identified by predefined distribution-based method[20] (left) and ClusterMap (right), respectively. **b,c**, As in (**a)** but using the STARmap[6] mouse placenta 903-gene dataset and published pciSeq[4] dataset. **d**, Bar plots demonstrating the remaining RNA numbers, cell numbers and segmentation accuracy for each dataset. In each bar plot, results from predefined distribution-based method[20], ClusterMap, and published reports were shown in blue, red, and yellow, respectively.

Supplementary Table

Supplementary Table 1. Summary of the name, *in situ* sequencing protocol, number of genes, number of cells, number of reads, number of cell types, corresponding figures and note of 7 datasets.

Dataset	Experimental Method	# Genes	# Cells	# Reads	# Cell types	Figures	Note
STARmap mouse V1 1020-gene	STARmap	1,020	1,599	863,426	16	Fig. 1c, Fig. 2, Supplementary Figs. 3,4,5	Source: Ref. 6. 2D analysis
STARmap mouse placenta 903-gene	STARmap	903	7,224	5,090,930	12	Fig. 3, Fig. 4, Supplementary Figs. 6,7	New data. 2D analysis
MERFISH mouse POA	MERFISH	140	10,320	3,065,171	9	Fig. 5, Supplementary Fig. 8	Source: Ref. 3. 3D analysis
pciSeq mouse hippocampus	ISS	98	2,924	31,750	23	Fig. 5, Supplementary Fig. 9	Source: Ref. 4. 2D analysis
osmFISH mouse SSp	osmFISH	33	8,538	1,690,328	31	Fig. 5, Supplementary Fig. 8	Source: Ref. 5. 2D analysis
STARmap cardiac organoid 8-gene	STARmap	8	1,519	47,594	3	Fig. 6, Supplementary Fig. 10	New data, 3D analysis
STARmap mouse V1 28-gene	STARmap	28	24,590	753,396	11	Fig. 6, Supplementary Fig. 10	Source: Ref. 6. 3D analysis