Cluster-guided contrastive learning with masked autoencoder for spatial domain identification based on spatial transcriptomics

Juan Wang[1,2], Qi Gao[1], Shasha Yuan[1], and Junliang Shang[1]

1. School of Computer Science, Qufu Normal University, Rizhao 276826, China

2. RizhaoQufu Normal University Joint Technology Transfer Center, Qufu Normal University, Rizhao 276826, China

A. Supplementary Table

TABLE S1

ARI scores from STMCCL and nine comparison methods [1-9] across 12 DLPFC sections.

#slice	STMCCL	STAGATE	SEDR	GraphST	DeepST	CCST	SpaGCN	SCANPY	SpaceFlow	stLearn
#151507	0.5742	0.6011	0.5223	0.4365	0.4707	0.511	0.449	0.3503	0.4391	0.323
#151508	0.5796	0.5402	0.4866	0.4541	0.4221	0.4584	0.354	0.3048	0.3809	0.3125
#151509	0.5928	0.4857	0.4168	0.4699	0.4768	0.4152	0.481	0.3911	0.3309	0.3173
#151510	0.5682	0.4654	0.5261	0.5137	0.4819	0.4594	0.404	0.3457	0.3043	0.2228
#151669	0.6996	0.3607	0.4263	0.4496	0.3488	0.4036	0.157	0.2439	0.3243	0.208
#151670	0.7004	0.3995	0.3295	0.3865	0.3578	0.346	0.359	-------	0.2019	0.1511
#151671	0.8252	0.6013	0.5649	0.639	0.4964	0.5976	0.483	0.4745	0.331	0.2442
#151672	0.789	0.5903	0.5839	0.6254	0.4936	0.548	0.564	0.3657	0.3398	0.2041
#151673	0.6697	0.5841	0.5895	0.5469	0.5836	0.5129	0.461	0.3297	0.4049	0.2152
#151674	0.6083	0.4979	0.6177	0.364	0.5438	0.5707	0.323	0.3219	0.3029	0.1891
#151675	0.6134	0.5948	0.6132	0.525	0.4892	0.5129	0.376	0.3634	0.3437	0.2527
#151676	0.606	0.6143	0.5698	0.5053	0.5263	0.5062	0.3	0.3193	0.3166	0.1759

TABLE S2

ARI scores from STMCCL and nine comparison methods across 12 DLPFC sections.

#slice	STMCCL	STAGATE	SEDR	GraphST	DeepST	CCST	SpaGCN	SCANPY	SpaceFlow	stLearn
#151507	0.6988	0.7058	0.6611	0.6508	0.663	0.6132	0.56	0.4534	0.5413	0.426
#151508	0.6386	0.677	0.6478	0.6279	0.6546	0.5745	0.459	0.4109	0.5253	0.4125
#151509	0.6803	0.6495	0.6159	0.6488	0.6228	0.6266	0.607	0.4858	0.5056	0.4902
#151510	0.672	0.6354	0.6636	0.6446	0.6227	0.601	0.535	0.4126	0.4753	0.3972
#151669	0.6841	0.5811	0.581	0.558	0.5533	0.5468	0.349	0.3708	0.4588	0.2973
#151670	0.629	0.5742	0.5014	0.5031	0.5362	0.4966	0.48	------	0.3656	0.2533
#151671	0.75	0.7015	0.6749	0.7327	0.7098	0.6301	0.601	0.5094	0.4227	0.3519

#151672	0.7629	0.6866	0.6867	0.7211	0.6669	0.6111	0.658	0.4391	0.4751	0.3101
#151673	0.7214	0.7175	0.712	0.6724	0.6817	0.6933	0.625	0.4805	0.5379	0.3922
#151674	0.6975	0.6726	0.7195	0.5306	0.6797	0.6491	0.489	0.3985	0.4202	0.3277
#151675	0.7275	0.7042	0.7176	0.638	0.6777	0.6599	0.522	0.4417	0.4843	0.4176
#151676	0.7121	0.7265	0.6802	0.6597	0.6217	0.6761	0.501	0.4164	0.4393	0.3259

B. clustering evaluation metrics

For different datasets, we used various clustering evaluation metrics to assess the expressiveness of the low-dimensional embeddings extracted by STMCCL. The ARI is formulated as:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]},$$

where the unadjusted rand index (RI) is defined as $RI = (a + b)/C_n^2$, with a being the number of pairs correctly labeled as coming from the same set, b being the number of pairs correctly labeled as not in the same set, and C_n^2 being the total number of possible pairs. $E[RI]$ is the expected RI of random labeling.

Mutual information (MI) measures the similarity between ground truth and predicted clusters. It is defined as:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

where $|U_i|$ is the number of the samples in cluster U_i and $|V_j|$ is the number of the samples in cluster $|V_j| \cdot MI$ is generally higher for clustering results with larger number of clusters. To account this bias, the Normalized Mutual Information (NMI) was calculated to remove the effect of cluster numbers:

$$NMI(U,V) = \frac{MI(U,V)}{F(H(U),H(V))}$$

where F can find functions of maximum, minimum, geometric mean and arithmetic mean.

F1 Score is an indicator used in statistics to measure the accuracy of the model. It is denoted as:

$$F1 = 2\frac{P \cdot R}{P+R}$$

where P represents precision, which measures the proportion of correctly predicted positive samples among all samples classified as positive. Similarly, R denotes recall, also known as the retrieval rate, which quantifies the proportion of correctly identified positive samples among all actual positive samples.

Finally, the formulas for SC and DB are as follows:

$$SC: \quad s(i) = \frac{a(i)-b(i)}{max(a(i),b(i))}$$

$$DB: \quad R_{pq} = 2\frac{S_p+S_q}{d_{pq}}$$

where $a(i)$ represents the average distance from point i to other points within the same cluster, and $b(i)$ represents the average distance from i to all points in the nearest cluster. Additionally, S_p and S_q are the average intra-cluster distances for cluster p and cluster q, respectively, while d_{pq} is the distance between cluster p and cluster q. The DB index is the average of the maximum R_{pq} value across all cluster pairs.

C. Visualization

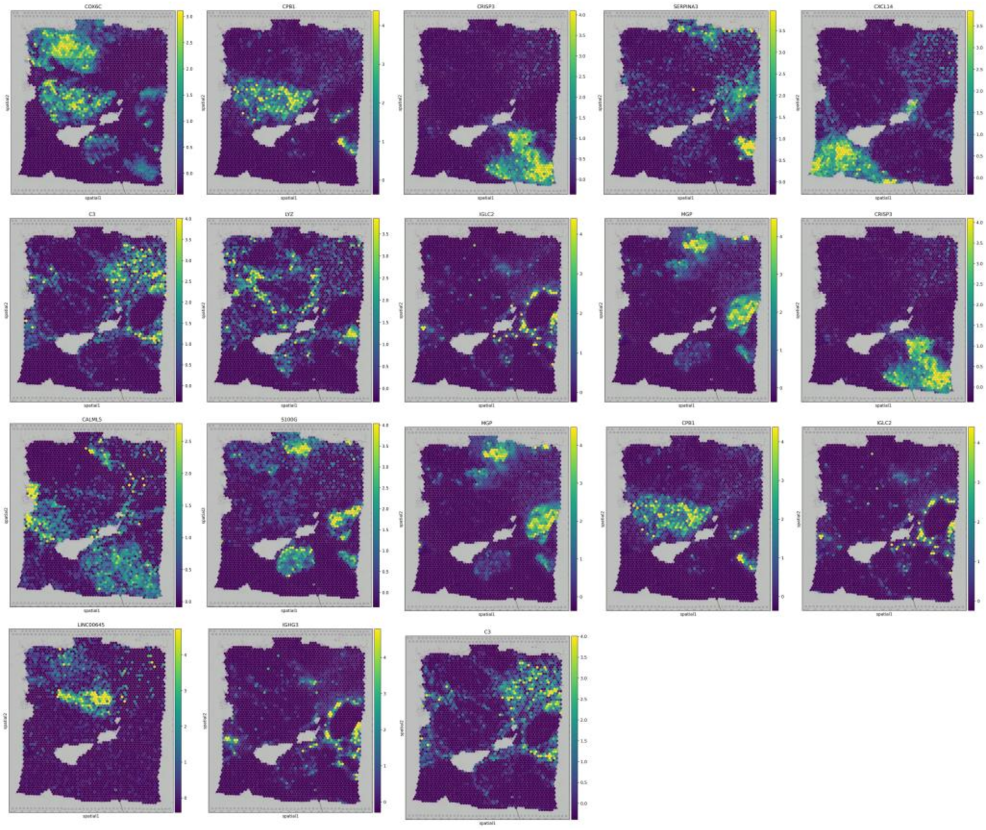

Fig. S1. Other high-expression genes in domains identified by STMCCL on the HBC dataset.

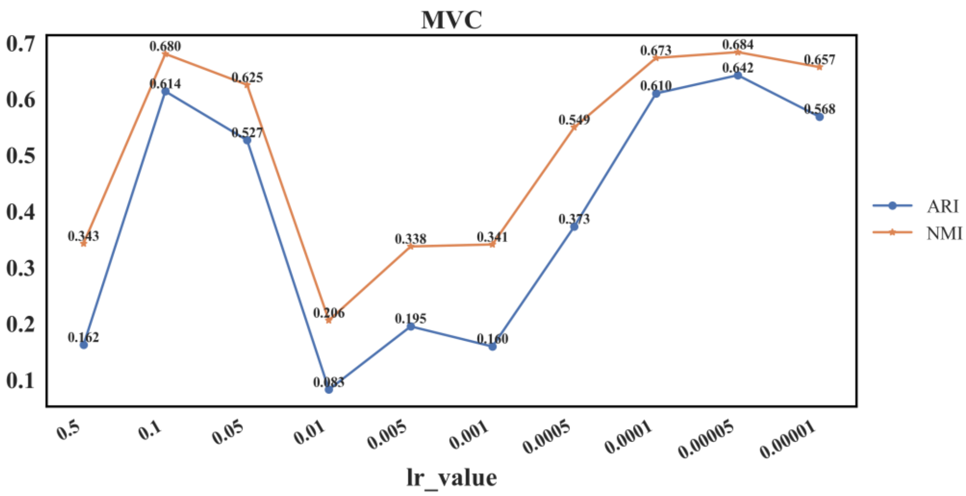

Fig. S2. The impact of learning rate on the MVC dataset.

D. References

[1] K. Dong and S. Zhang, "Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder," Nature communications, vol. 13, no. 1, p. 1739, 2022.

[2] H. Xu, H. Fu, Y. Long, K. S. Ang, R. Sethi, K. Chong, M. Li, R. Uddamvathanak, H. K. Lee, J. Ling et al., "Unsupervised spatially embedded deep representation of spatial transcriptomics," Genome Medicine, vol. 16, no. 1, p. 12, 2024.

[3] Y. Long, K. S. Ang, M. Li, K. L. K. Chong, R. Sethi, C. Zhong, H. Xu, Z. Ong, K. Sachaphibulkij, A. Chen et al., "Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with graphst," Nature Communications, vol. 14, no. 1, p. 1155, 2023.

[4] C. Xu, X. Jin, S. Wei, P. Wang, M. Luo, Z. Xu, W. Yang, Y. Cai, L. Xiao, X. Lin et al., "Deepst: identifying spatial domains in spatial transcriptomics by deep learning," Nucleic Acids Research, vol. 50, no. 22, pp. e131–e131, 2022.

[5] J. Li, S. Chen, X. Pan, Y. Yuan, and H.-B. Shen, "Cell clustering for spatial transcriptomics data with graph neural networks," Nature Computational Science, vol. 2, no. 6, pp. 399–408, 2022.

[6] J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li, "Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network," Nature methods, vol. 18, no. 11, pp. 1342–1351, 2021.

[7] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," Genome biology, vol. 19, pp. 1–5, 2018.

[8] H. Ren, B. L. Walker, Z. Cang, and Q. Nie, "Identifying multicellular spatiotemporal organization of cells with spaceflow," Nature communications, vol. 13, no. 1, p. 4076, 2022.

[9] D. Pham, X. Tan, J. Xu, L. F. Grice, P. Y. Lam, A. Raghubar, J. Vukovic, M. J. Ruitenberg, and Q. Nguyen, "stlearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues," BioRxiv, pp. 2020–05, 2020.