Supplementary Material - Refinement Contrastive Learning of Cell-Gene Associations for Unsupervised Cell Type Identification

A Introduction

This supplementary material aims to offer additional technical details and supporting evidence to facilitate a deeper understanding of our proposed scRCL. Specifically, this document details the overall model optimization process, provides comprehensive information on the single-cell omics datasets (including both single-cell transcriptomics and spatial transcriptomics datasets) used in the experiments, and offers additional experiments to further support the evaluation of this approach. These elaborations contribute to a clearer understanding of the proposed scRCL.

B Detailed Model Optimization

While the main text provides an in-depth explanation of the individual components of our proposed method, scRCL, we present the complete model optimization procedure in Algorithm 1 for clarity and reproducibility. The training process consists of the following steps:

- We first construct the cell KNN graph $\mathbf{A}$ from the gene expression matrix in scRNA-seq data, or from spatial coordinates in spatial transcriptomics data. Simultaneously, we build the gene graph $\mathbf{G}$ using $\mathbf{X}^\top$.
- We then obtain heterogeneous latent embeddings $\mathbf{E}^m$ and $\mathbf{E}^g$ by passing $\mathbf{X}$ through the MLP networks ϕ and GNN networks ψ, respectively, as defined in Equation (1). The latent gene embeddings $\mathbf{U}$ are obtained by encoding $\mathbf{X}^\top$ together with the gene graph $\mathbf{G}$ using the gene networks φ, following Equation (6).
- Based on the latent embedding distributions $\mathbf{E}^m$ and $\mathbf{E}^g$, we calculate $\mathcal{L}_{HEA}$ (Equation (3)) to achieve both global and cell-level distribution alignment. In parallel, we calculate the instance-level pairwise cross-view symmetric KL divergence matrix κ, which is used to compute the neighborhood distribution contrastive alignment loss $\mathcal{L}_{NDC}$ (Equation (4)), thereby aligning cellular neighborhood distribution across views.
- Next, we refine the latent embeddings using the cell-gene association representation $\mathbf{U}$ to produce enhanced cell representations $\mathbf{Z}^m$ and $\mathbf{Z}^g$ via Equation (7). From these, we derive the cross-view cosine similarity matrix $\mathbf{S}$ and compute the cross-view correlation contrastive loss $\mathcal{L}_{CVC}$, which strengthens consistency and structural fidelity of the learned representations through cross-view correlation contrastive learning.
- Subsequently, the entire model is Optimized by jointly minimizing $\mathcal{L}_{HEA}$, $\mathcal{L}_{NDC}$ and $\mathcal{L}_{CVC}$. The above procedure is repeated for I iterations until convergence.
- Finally, we apply k-means algorithm to the concatenated representation $\mathbf{Z}$ (Equation (11)) for each cell to produce cluster assignments, enabling a wide range of meaningful downstream analyses.

This step-by-step workflow aims to offer readers a more comprehensive understanding of how the overall model is trained and optimized in practice.

Algorithm 1: Optimization of scRCL

Input: Single cell omics data $\mathbf{X}$.
Parameter: Iteration number I; cluster number c; hyperparameters α, β and k.

```
Calculate cell KNN graph A and gene graph G.
for epoch = 1 to I do
    Obtain the heterogeneous latent embeddings E^m, E^g via Equation (1).
    Obtain U via Equation (6).
    Calculate L_HEA with Equation (3).
    Calculate κ via Equation (5), and L_NDC with Equation (4).
    Refine cell embeddings to obtain Z^m, Z^g via Equation (7).
    Calculate S based on Z^m, Z^g, and then compute L_CVC via Equation (9).
    Optimize all networks by jointly minimizing L_HEA, L_NDC and L_CVC.
end for
Perform the k-means algorithm on the final representation Z for each cell to produce cluster assignments.
```

Output: Clustering results.
Conduct diverse downstream analyses.

C Dataset Description

In this section, we provide detailed descriptions of all the single-cell transcriptomics and spatial transcriptomics

Dataset	Tissue	Platform	#Cells	#Populations	#Genes
Tumor	Human Colorectal Tumor	Illumina HiSeq	363	6	13495
Diaphragm	Mouse Diaphragm	Smart-seq2	870	5	23341
Lung	Mouse Lung	Smart-seq2	1676	11	23341
Trachea	Mouse Trachea	Smart-seq2	1350	4	23341
Human_ESC	Human Embryo	Illumina HiSeq	1018	7	18294
Zeisel	Mouse Brain	STRT-Seq UMI	3005	9	11392
Bladder	Mouse Bladder	10x	2500	4	23341
Limb_Muscle	Mouse Limb Muscle	10x	3909	6	23341
Spleen	Mouse Spleen	10x	9552	5	23341
Baron_Human	Human Pancreas	inDrop	8569	14	20125

Table 1: scRNA-seq Dataset Statistics.

Platform	Tissue	Section	#Spots/Bins	#Populations	#Genes
10x Visium	Human Dorsolateral Prefrontal Cortex (DLPFC)	151507	4221	7	33538
		151508	4381	7	33538
		151509	4788	7	33538
		151510	4595	7	33538
		151669	3636	5	33538
		151670	3484	5	33538
		151671	4093	5	33538
		151672	3888	5	33538
		151673	3611	7	33538
		151674	3635	7	33538
		151675	3566	7	33538
		151676	3431	7	33538
	Human Breast	Human Breast Cancer Section 1	3798	20	36601
	Mouse Brain	Mouse Brain Section 1 (Sagittal-Anterior)	2695	52	32285
Stereo-seq	Mouse Embryo	E9.5	5913	12	25568

Table 2: Spatial Transcriptomics Dataset Statistics.

datasets used in our experiments. single-cell transcriptomics datasets: Tumor(Li et al. 2017), Diaphragm, Lung, Trachea, Bladder, Limb_Muscle, Spleen (Schaum et al. 2018), Human_ESC(Chu et al. 2016), Zeisel(Zeisel et al. 2015), Baron_Human(Baron et al. 2016). Spatial transcriptomics datasets: LIBD human dorsolateral prefrontal cortex (DLPFC) (Maynard et al. 2021) with 12 tissue slices, Stereo-seq E9.5 Mouse Embryo (Chen et al. 2022), Human Breast Cancer (Buache et al. 2011), Mouse Brain Anterior (Lein et al. 2007).

Comprehensive dataset statistics are summarized in Table 1 (single-cell transcriptomics datasets) and Table 2 (spatial transcriptomics datasets). Specifically, we report for each dataset the associated tissue of origin, the sequencing platform used for data acquisition, the number of cells (or spots/bins), the number of annotated cell populations, and the number of genes utilized for downstream analysis. These datasets, which vary in terms of cell numbers, tissue types, and sequencing platforms, serve as a robust benchmark to comprehensively evaluate the effectiveness and generalizability of our proposed method.

D Implementation Details

Following previous works, we normalize total counts per cell and apply logarithmic transformation of the gene expression data for all datasets during pre-processing. Specifically, to filter out measurement outliers, we retain only cells with at least one gene expression count and genes expressed in at least one cell. To account for variations in count scales, we normalize the expression vector of each cell by the total number of genes, ensuring that the gene expression values are scaled to a comparable range across all cells.

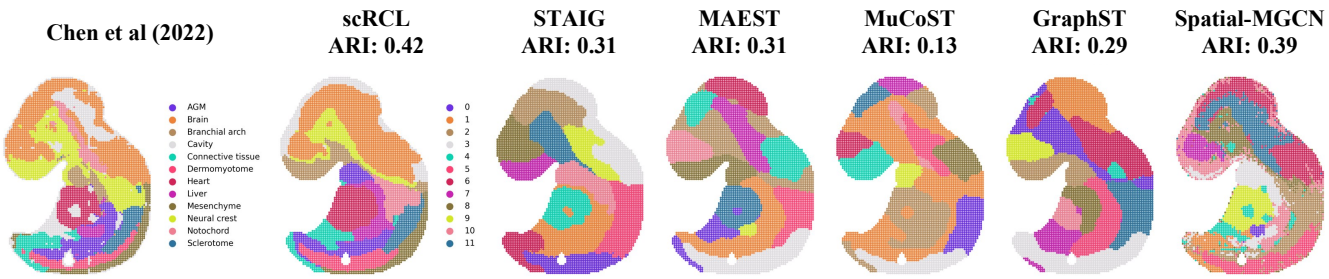

Figure 1: Visualization of spatial domain identification on the Stereo-seq Mouse Embryo E9.5 dataset.

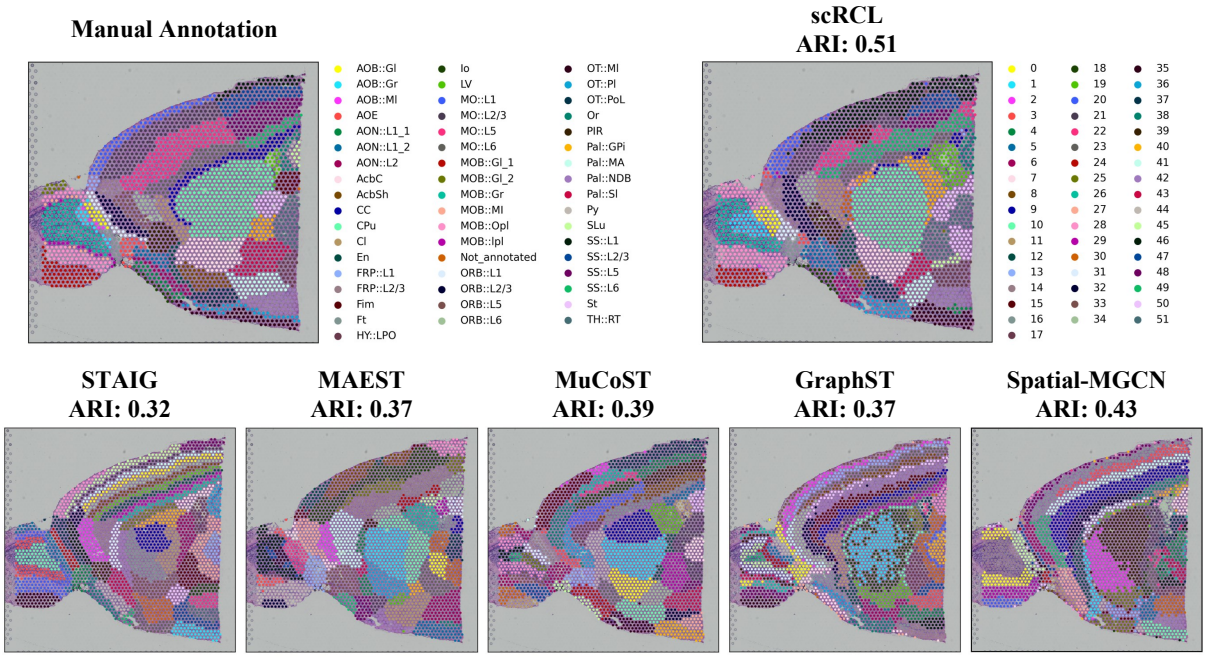

Figure 2: Visualization of spatial domain identification on the Mouse Brain Anterior dataset.

Subsequently, we apply a logarithmic transformation to the gene expression matrix to stabilize variance and select the top h highly-variable genes (HVGs) to form the final input features. All pre-processing steps are performed using the Scanpy toolkit.

All experiments are conducted on an Intel Xeon W-2235 CPU, and an NVIDIA 2080 Super GPU with the PyTorch 1.12.1 platform. The model is trained for 500 epochs until convergence using the Adam (Kingma and Ba 2017) optimizer. The initial learning rate is set as 1e-4 or 1e-5. The latent embeddings dimension d is fixed to 1500; the hyperparameters α, β are selected from $\{0.01, 0.1, 1.0, 10, 100, 200\}$; the numbers of nearest neighbors are set to 15 for the cell graph while 5 for the gene graph by default.

E Additional Experiments

To provide a comprehensive evaluation of our proposed scRCL, we conduct a series of additional experiments. These include the visualization of spatial domain identification on three spatial transcriptomics datasets, robustness analysis on highly variable genes, UMAP visualization analysis, and hyperparameter sensitivity analysis.

E.1 Visualization of Spatial Domain Identification on Three Spatial Transcriptomics Datasets

As reported in Table 2 of the main manuscript, scRCL achieves the highest ACC, NMI, and ARI scores among five competing methods across three spatial transcriptomics datasets. To further demonstrate the effectiveness of spatial domain identification, we visualize the predicted spatial domains of all methods in Figure 1 (Stereo-seq Mouse Embryo E9.5), Figure 2 (Mouse Brain Anterior), and Figure 3 (Human Breast Cancer).

It is evident from the visualizations that scRCL produces domain segmentation results that are most consistent with manual annotations. Specifically, on the Mouse Embryo E9.5 dataset, scRCL successfully identifies major anatomical regions such as brain, neural crest, mesenchyme, dermomyotome, sclerotome, heart, AGM, and liver. In contrast,

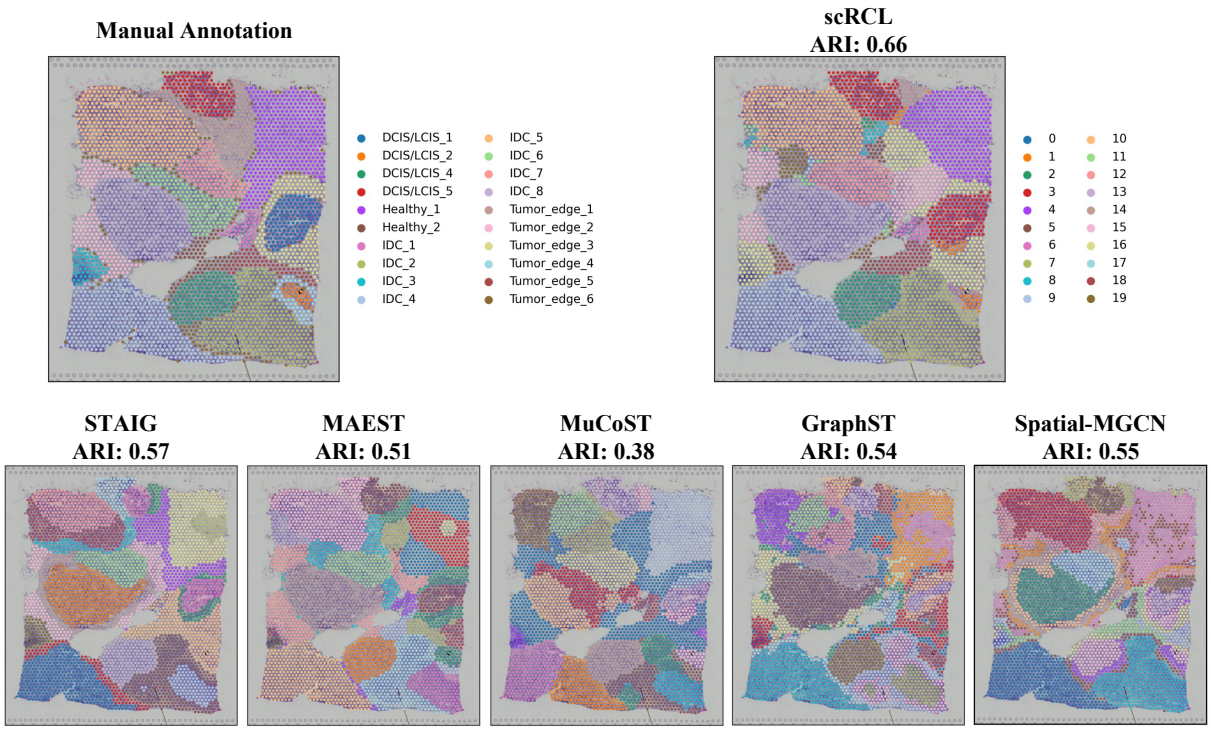

Figure 3: Visualization of spatial domain identification on the Human Breast Cancer dataset.

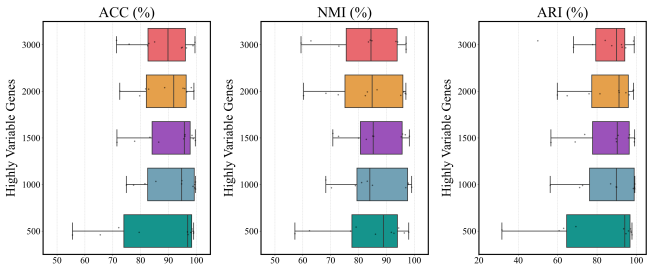

Figure 4: Clustering results under different number of highly variable genes selections on ten scRNA-seq datasets

other methods exhibit noticeable confusion among mesenchyme, dermomyotome, and sclerotome.

On the Mouse Brain Anterior dataset, scRCL accurately distinguishes spatial domains including MOB::Gr, MOB::Gl_1, MOB::Opl, AOB::Gl, CPu, AcbC, and Pal::NDB, with particularly clear boundaries among MOB::Gr, MOB::Gl_1, and MOB::Opl. Other methods fail to clearly identify critical regions such as CPu and AcbC. For instance, STAIG and MAEST over-segment CPu into three different regions, MuCoST and Spatial-MGCN split it into two, while GraphST produces fragmented and disorganized clusters.

On the Human Breast Cancer dataset, scRCL demonstrates a clear advantage over baseline methods with respect to ARI score. It correctly recovers most of the regions including IDC_1 to IDC_5, IDC_8, DCIS/LCIS_1, DCIS/LCIS_2, DCIS/LCIS_4, DCIS/LCIS_5, Healthy_1, and partial regions of Tumor_edge_2, Tumor_edge_3, Tumor_edge_5, and Tumor_edge_6. Other methods fail to accurately identify key areas such as IDC_5, DCIS/LCIS_1, and DCIS/LCIS_5.

In summary, scRCL consistently provides more accurate and biologically meaningful spatial domain identification across diverse datasets, outperforming existing baseline methods in both quantitative metrics and visual interpretability.

E.2 Robustness Analysis

In single-cell omics data analysis, a common pre-processing step is to select highly variable genes (HVGs) to filter out uninformative genes. The number of HVGs selected can significantly influence the model's performance. To assess the robustness of scRCL to different HVG selection settings, we conduct experiments on 10 scRNA-seq datasets by varying the number of selected HVGs from the set $\{500, 1000, 1500, 2000, 3000\}$. The results are shown in Figure 4.

We observe that scRCL achieves consistently stable clustering performance when the number of HVGs ranges from 1000 to 2000, with optimal performance typically observed at 1000. In contrast, performance degrades noticeably when only 500 HVGs are used, likely due to reduced representation capacity caused by the limited input information. Overall, these results demonstrate that scRCL exhibits strong robustness within a reasonable range of HVG selection, particularly between 1000 and 2000 genes.

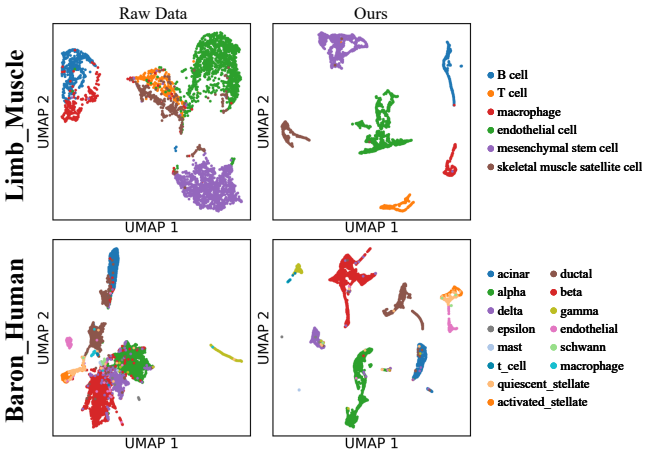

Figure 5: UMAP visualization on two scRNA-seq datasets.

E.3 UMAP Visualization

To further evaluate the quality of the learned cell representations, we apply UMAP (McInnes, Healy, and Melville 2018) to visualize the embeddings obtained by scRCL, as shown in Figure 5. The left panel displays the UMAP projection based on the original gene expression data, while the right panel shows the projection based on the embeddings learned by our model. As illustrated, the representations produced by scRCL exhibit clearer cluster boundaries and are more discriminative compared to the original input features. This demonstrates that our model successfully captures higher-quality representations, which contribute to improved clustering performance and better support for various downstream analyses.

E.4 Hyperparameter Analysis

Loss Weighting Coefficients α and β. In Equation (10) of the main text, α and β denote the loss weighting coefficients for $\mathcal{L}_{NDC}$ and $\mathcal{L}_{CVC}$, respectively. To investigate the sensitivity of model performance to these two hyperparameters, we conduct experiments on two scRNA-seq datasets by varying α and β across the range $\{0.01, 0.1, 1.0, 10.0, 100.0\}$. The clustering metrics (ACC, NMI, ARI) under different settings are reported in Figure 6. From the results, we observe that when α is very small, the model fails to achieve satisfactory performance. In contrast, when $\alpha \in [10, 100]$ and $\beta \in [0.01, 1.0]$, the model consistently achieves strong clustering results. These findings suggest that scRCL is robust within a relatively broad and practical range of hyperparameter choices for α and β.

The Number of Nearest Neighbors k. When constructing the cell graph, we select the k nearest neighbors as the spatial neighbors for each cell. The choice of k can significantly influence model performance. To evaluate the sensitivity of scRCL to this parameter, we conduct experiments on all scRNA-seq datasets by varying k, as illustrated in Figure 7. The results clearly show that scRCL achieves optimal performance on most datasets when k is within the range [10, 15]. This can be attributed to the fact that a small k may

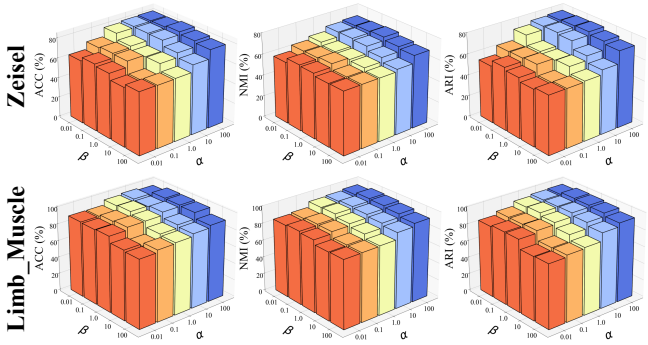

Figure 6: Clustering results with different α and β on two datasets.

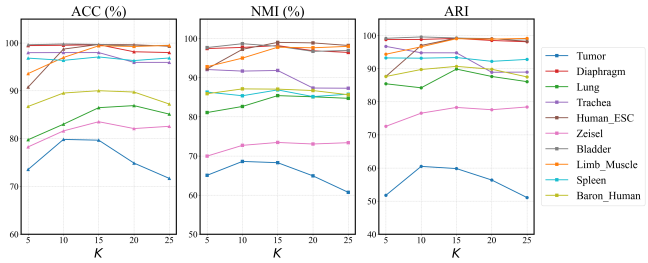

Figure 7: Clustering results with different number of nearest neighbors k on ten datasets.

fail to capture sufficient semantic context, while a large k may introduce noisy or semantically inconsistent neighbors, which can negatively affect representation learning. Based on this analysis, we set the default value of k to 15 in our experiments, which consistently leads to superior performance compared to baseline methods.

References

Baron, M.; Veres, A.; Wolock, S. L.; Faust, A. L.; Gaujoux, R.; Vetere, A.; Ryu, J. H.; Wagner, B. K.; Shen-Orr, S. S.; Klein, A. M.; et al. 2016. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems*, 3(4): 346–360.

Buache, E.; Etique, N.; Alpy, F.; Stoll, I.; Muckensturm, M.; Reina-San-Martin, B.; Chenard, M.; Tomasetto, C.; and Rio, M. 2011. Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice. *Oncogene*, 30(29): 3261–3273.

Chen, A.; Liao, S.; Cheng, M.; Ma, K.; Wu, L.; Lai, Y.; Qiu, X.; Yang, J.; Xu, J.; Hao, S.; et al. 2022. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10): 1777–1792.

Chu, L.-F.; Leng, N.; Zhang, J.; Hou, Z.; Mamott, D.; Vereide, D. T.; Choi, J.; Kendziorski, C.; Stewart, R.; and Thomson, J. A. 2016. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology*, 17(1): 173.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lein, E. S.; Hawrylycz, M. J.; Ao, N.; Ayres, M.; Bensinger, A.; Bernard, A.; Boe, A. F.; Boguski, M. S.; Brockway, K. S.; Byrnes, E. J.; et al. 2007. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124): 168–176.

Li, H.; Courtois, E. T.; Sengupta, D.; Tan, Y.; Chen, K. H.; Goh, J. J. L.; Kong, S. L.; Chua, C.; Hon, L. K.; Tan, W. S.; et al. 2017. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5): 708–718.

Maynard, K. R.; Collado-Torres, L.; Weber, L. M.; Uytingco, C.; Barry, B. K.; Williams, S. R.; Catallini, J. L.; Tran, M. N.; Besich, Z.; Tippani, M.; et al. 2021. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3): 425–436.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

Schaum, N.; Karkanias, J.; Neff, N. F.; May, A. P.; Quake, S. R.; Wyss-Coray, T.; Darmanis, S.; Batson, J.; Botvinnik, O.; Chen, M. B.; et al. 2018. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium. *Nature*, 562(7727): 367.

Zeisel, A.; Muñoz-Manchado, A. B.; Codeluppi, S.; Lönnerberg, P.; La Manno, G.; Juréus, A.; Marques, S.; Munguba, H.; He, L.; Betsholtz, C.; et al. 2015. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226): 1138–1142.