Supplementary Methods

Simulation of counts

We first introduce how the true cell types were determined in our simulation. For Datasets 1-3, cell type labels were not available in the original publications, so we applied the consensus clustering method SC3 [1] to obtain cell type labels. SC3 uses a consensus approach to obtain robust clustering assignments, and it has been shown to have favorable results in two independent benchmarking studies [2, 3]. For Datasets 4 and 7, previously annotated cell type labels were directly used [4, 5]. For Dataset 5 [6], we selected the most abundant cell sub-type in each major cell type reported in the original publication. For Dataset 6 [7], we selected the eight most abundant cell types reported in the original publication.

Then, for each dataset, we assigned cell type labels to the individual cells (or spots) using a custom RShiny application. In this assignment process, we took the real H&E-stained images and spatial patterns of cell type labels in real data as references. For example, Supplementary Figure S1 provides an illustration of the R shiny app program used to assign true cell type labels based on spatial coordinates in Dataset 1. For each cell type label, we can use the application to manually select spots that belong to this cell type. Cells that have been assigned labels will be shown in the corresponding color while cells not assigned will remain in gray. We repeat this process for every cell type until all the spots have been assigned.

For each cell type, we then used scDesign2 to learn gene expression parameters from corresponding real data and to generate simulated read counts for the number of synthetic cells (or spots) determined through the RShiny application. The counts used for the five technical replicates of Datasets 2, 5, 6, and 7 in the robustness to sequencing depth section were generated by changing the random seed used for scDesign2.

Simulation of H&E images

For each real dataset, the color range was defined by selecting a real H&E-stained image and choosing one pixel from the darkest region with red, green, and blue values denoted as r_d, g_d, and b_d and one pixel from the lightest region with values r_l, g_l, and b_l. Ratios of red to green (R/G: $r_d/g_d, r_l/g_l$) and red to blue (R/B: $r_d/b_d, r_l/b_l$) were then calculated for both pixels and used to build two uniform distributions of the R/G and R/B ratios. For every cell (or spot) in the simulated dataset, its corresponding R/G and R/B ratios in the simulated H&E-stained image were randomly sampled from these uniform distributions. In addition, a normal distribution (truncated at r_d and r_l) was used to generate the red channel values. Given a total of K true cell types, for cells (spots) in cell type k, the mean of the untruncated normal distribution is set to $\mu + \frac{\lambda_k |r_l - r_d|}{C}$, where $\mu = \frac{r_l + r_d}{2}$ and $(\lambda_1, \lambda_2, \ldots, \lambda_K)$ is a random permutation of (0,1,...,K-1). The standard deviation of the untruncated distribution is set to 0.5. Simulated red channel values were then divided by previously sampled R/G and R/B ratios to produce the resulting green and blue channel values.

The simulated RGB values were plotted with cell/spot coordinates to produce final images for simulated data.

Generating H&E-Stained images with different variability

To generate H&E-Stained images with different levels of variation, we followed the simulation procedure described above but changed the standard deviation of the normal distributions. With a larger standard deviation, the generated RGB values for pixels across cell types become more similar, and the histology image has less useful information for the clustering of cell types. For every dataset, five values of the standard deviation were considered: 0.5, 10, 20, 50, and 100. In the clustering analysis, the spatial locations and gene expression levels were kept the same regardless of the histology images being used.

Evaluating methods with different cluster numbers as the input

Some clustering methods require or optionally allow a parameter input to specify the number of clusters. In order to better evaluate the clustering performance of each method in the context of unknown true cell type numbers, we tested each method in our analysis with varying parameter values. For each method that uses a parameter indicating the initial cluster number, we applied the method with five values of the parameter: $k-2, k-1, k, k+1, k+2$, where k is the true cell type number.

Software implementation details

All software parameters are assumed to be the default that is provided by the package of the specified version, and all pre-processing steps provided by the official repositories or vignettes of these packages are used unless otherwise specified. Any deviations to the default pipeline for any of the clustering methods are explained in this section. The seed `2020` was used for all random seed parameters among all methods.

BayesSpace version 1.00 was used, and functions `qTune` and `qPlot` were skipped as they required manual tuning. The `q` parameter of the `spatialCluster` function was set to the true cell type number.

DR.SC version 2.9 was used, and the `DR.SC` function with `maxIter` set to 20, the `K` parameter was set to the true cell type number, and the platform specified as `ST` was used for all DRSC-based clustering.

Giotto version 1.0.3 was used for all Giotto-based methods. `min_det_genes_per_cell` was set to 1 when creating the Giotto object. The `k` parameter of `doKmeans` and `center` parameter of `doHclust` were set to the true cell type number input for Giotto-KM and Giotto-H, respectively. The top 1500 (or number of genes returned by `binSpect` if it is greater than 1500) genes and the target ranges recommended by https://qzhudfci.bitbucket.io/spatialgiotto/giotto_ spatial_pattern_mining.html for the beta parameter were used for Giotto-HM. The maximum ARI across all beta parameters used by doHMRF was reported as the final ARI for Giotto-HM in the results section.

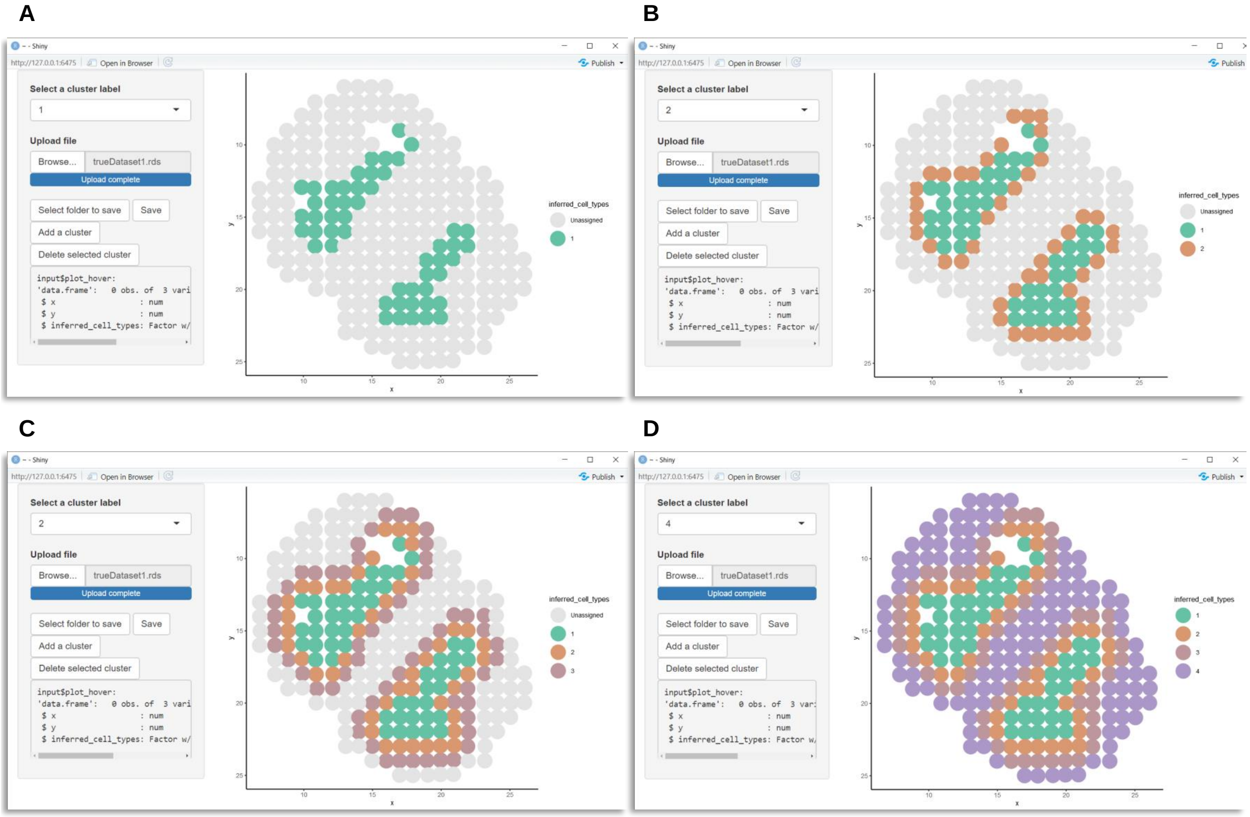
Seurat version 4.0.5 was used for Seurat-LV, Seurat-LVM, and Seurat-SLM. All Seurat-based clustering was performed using the FindClusters method with the resolution set as 0.5.

SpaCell version 1.0.1 was used. The image_normalization.py file was used for preprocessing and tiling of the simulated histology images, and the spacell_clustering.py file was used for clustering with the k parameter set to the true cell type number and the -l parameter set to mean_squared_error. The default pre-trained convolutional network was used. Additionally, for cases where the true cell type number exceeds 10, the default number of unique color_map values in the config.py file is not sufficient and causes a runtime error. For these cases, we add additional unique hex colors to these color_map values.

SpaGCN version 1.2.0 was used, and the n_clusters of the detect_spatial_domains_ez_mode function was set to the cluster number input for SpaGCN+ and SpaGCN. For SpaGCN, only the spatial coordinates were provided as input to the calculate_adj_matrix function, and the histology parameter was set to false.

stLearn version 0.3.2 was used. The min_cells parameter for filter_genes was set to 3. 50 (or number of genes if it's smaller than 50) components were used for the run_pca function, and the parameter values for add.image were: library_id="Old_ST", quality = "hires", scale=1.0, and spot_diameter_fullres=50.

Supplementary Figures



Select a cluster label: 1

Upload file: trueDataset1.rds

```
input$plot_hover:
'data.frame':   0 obs. of  3 vari
 $ x                  : num
 $ y                  : num
 $ inferred_cell_types: Factor w/
```

Select a cluster label: 2

Upload file: trueDataset1.rds

```
input$plot_hover:
'data.frame':   0 obs. of  3 vari
 $ x                  : num
 $ y                  : num
 $ inferred_cell_types: Factor w/
```

Select a cluster label: 2

Upload file: trueDataset1.rds

```
input$plot_hover:
'data.frame':   0 obs. of  3 vari
 $ x                  : num
 $ y                  : num
 $ inferred_cell_types: Factor w/
```

Select a cluster label: 4

Upload file: trueDataset1.rds

```
input$plot_hover:
'data.frame':   0 obs. of  3 var:
 $ x                  : num
 $ y                  : num
 $ inferred_cell_types: Factor w.
```

Figure S1: An illustration of the Rshiny app program used to assign true cell type labels based on spatial coordinates in Dataset 1.

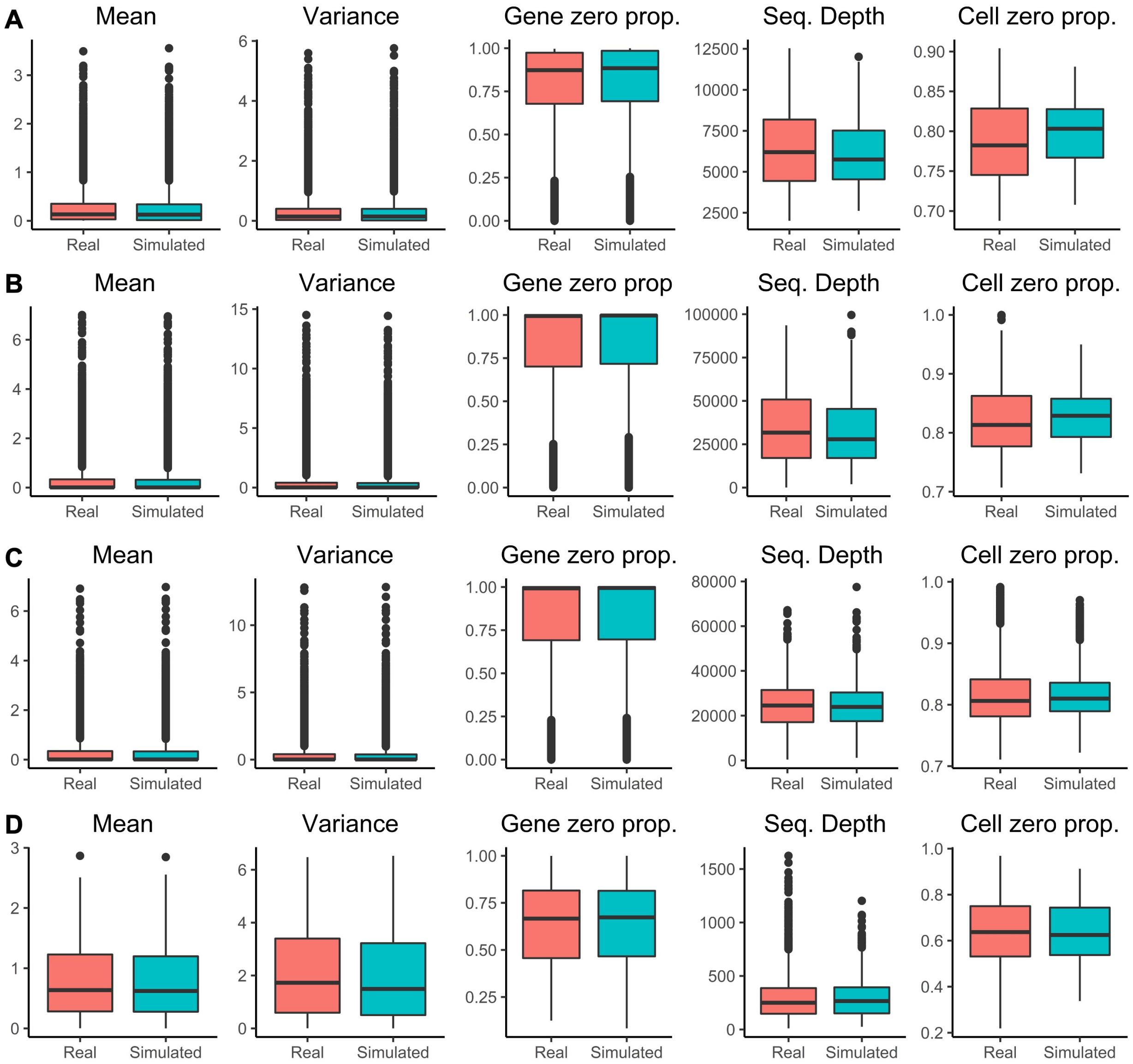


Figure S2: Comparison between simulated and real gene expression data for Datasets 1-4 (**A**-**D**). The count mean of a gene was calculated as log-transformed average counts. The count variance of a gene was calculated as log-transformed variance. The gene (cell) zero proportion was calculated as the proportion of zero counts across cells (genes).

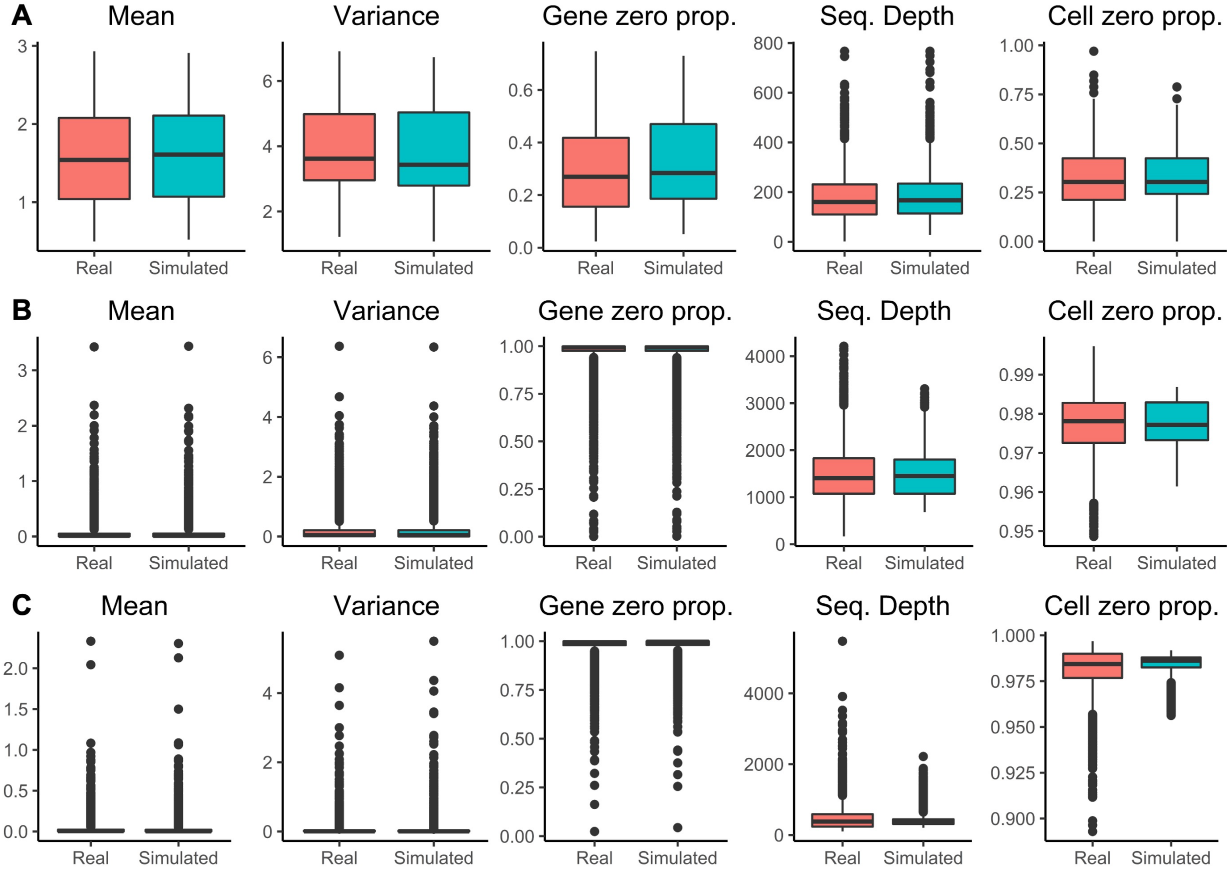


Figure S3: Comparison between simulated and real gene expression data for Datasets 5-7 (**A**-**C**). The count mean of a gene was calculated as log-transformed average counts. The count variance of a gene was calculated as log-transformed variance. The gene (cell) zero proportion was calculated as the proportion of zero counts across cells (genes).

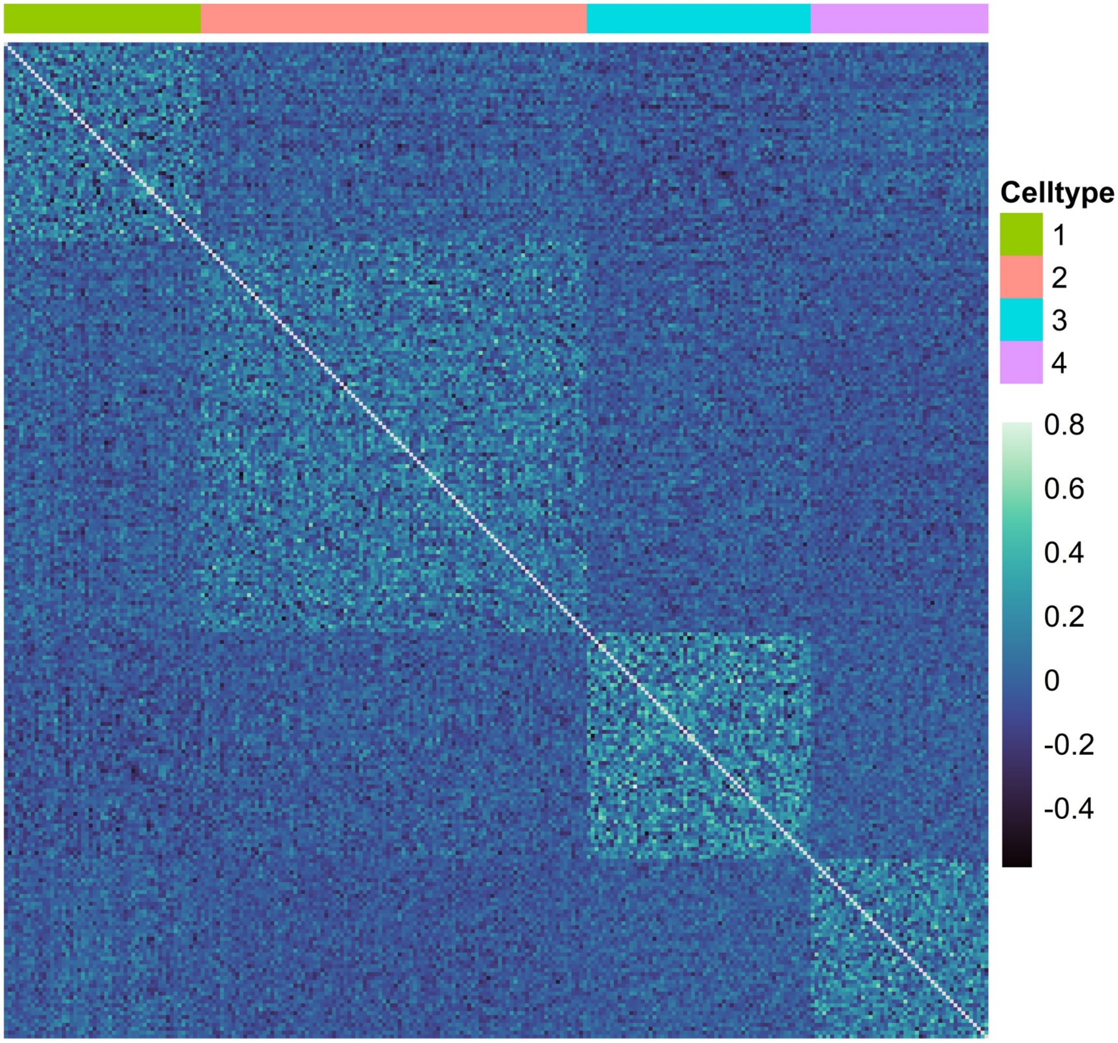


Figure S4: Cell-cell correlations in Dataset 1. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

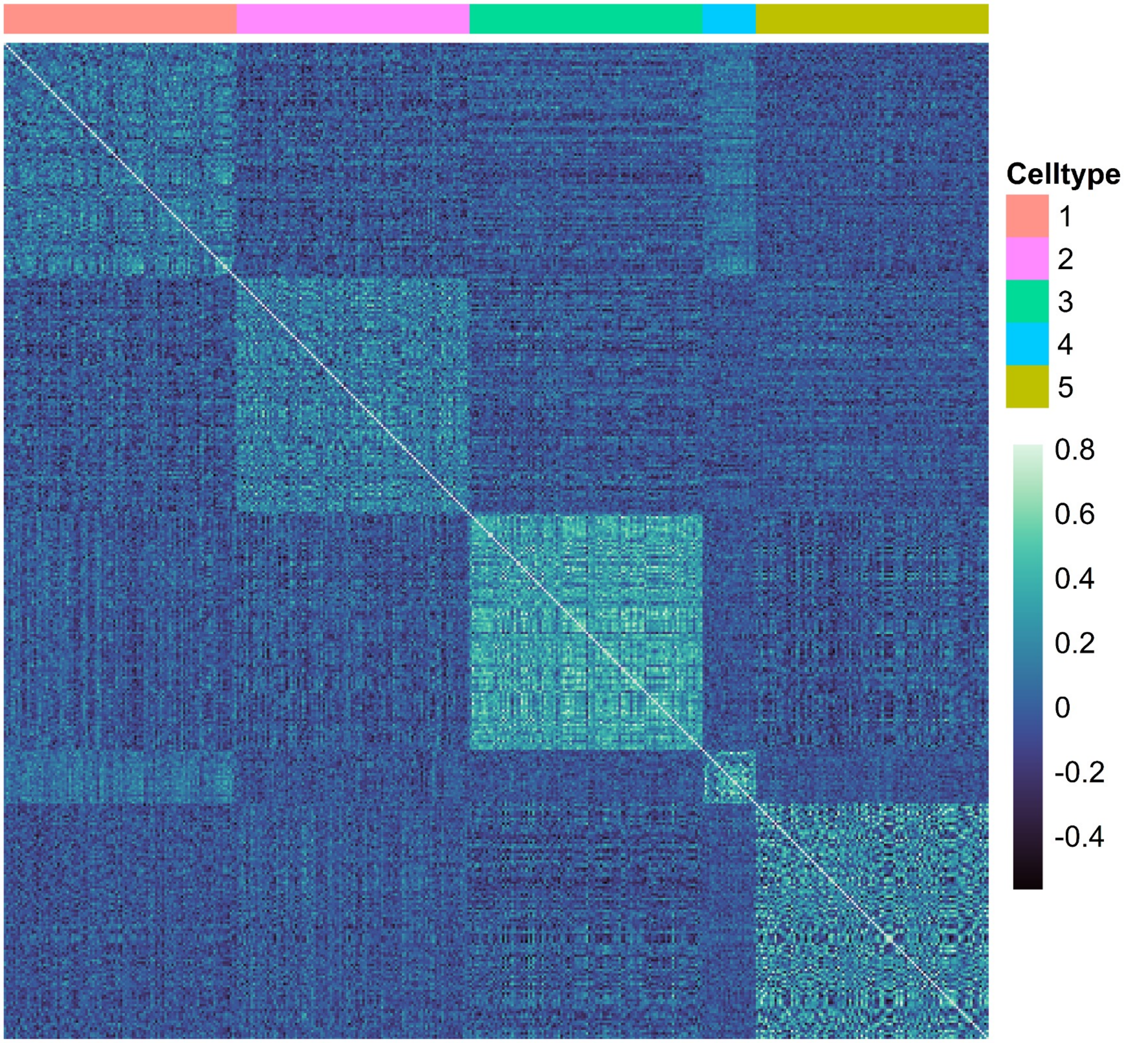


Figure S5: Cell-cell correlations in Dataset 2. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

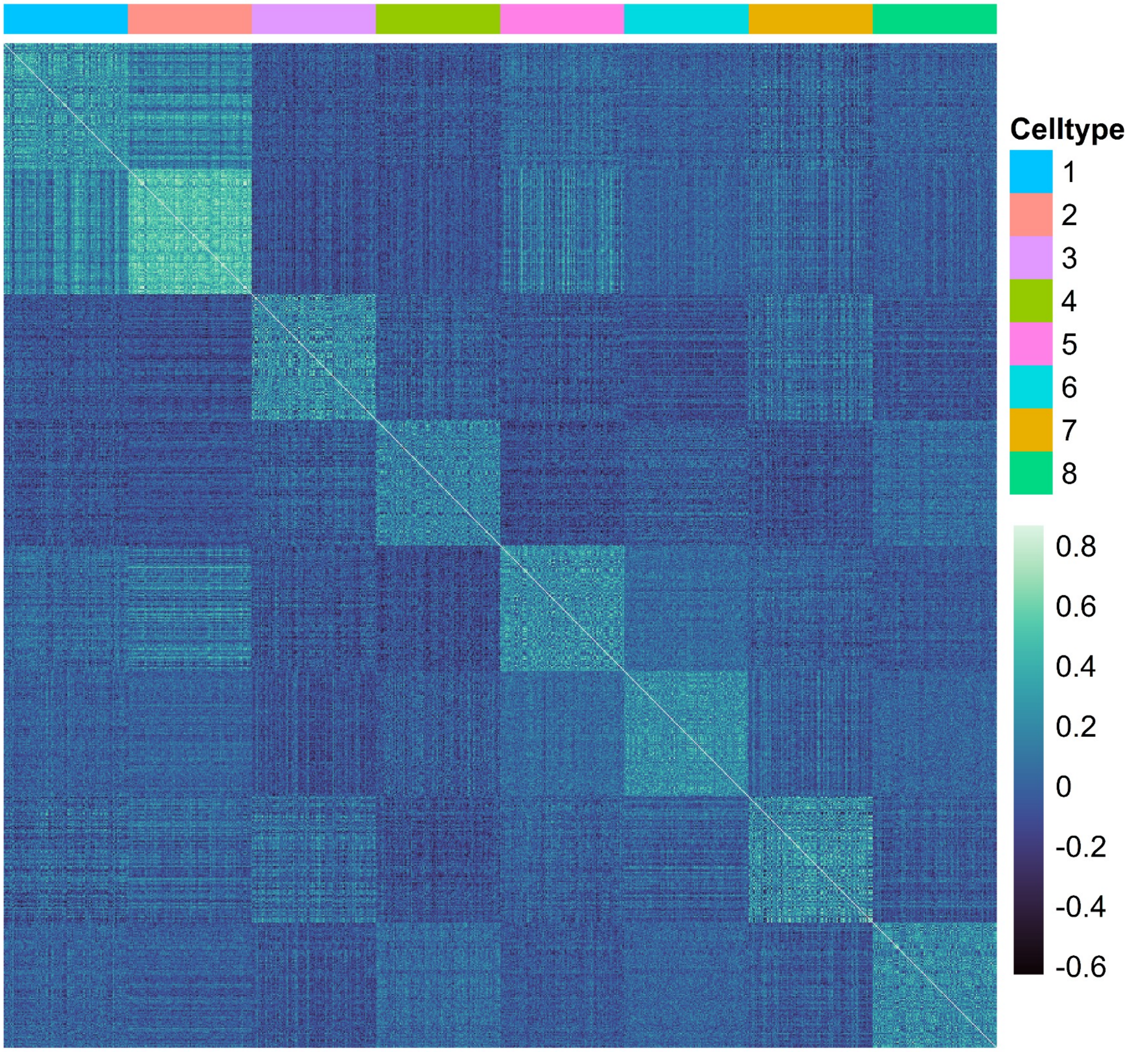


Figure S6: Cell-cell correlations in Dataset 3. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

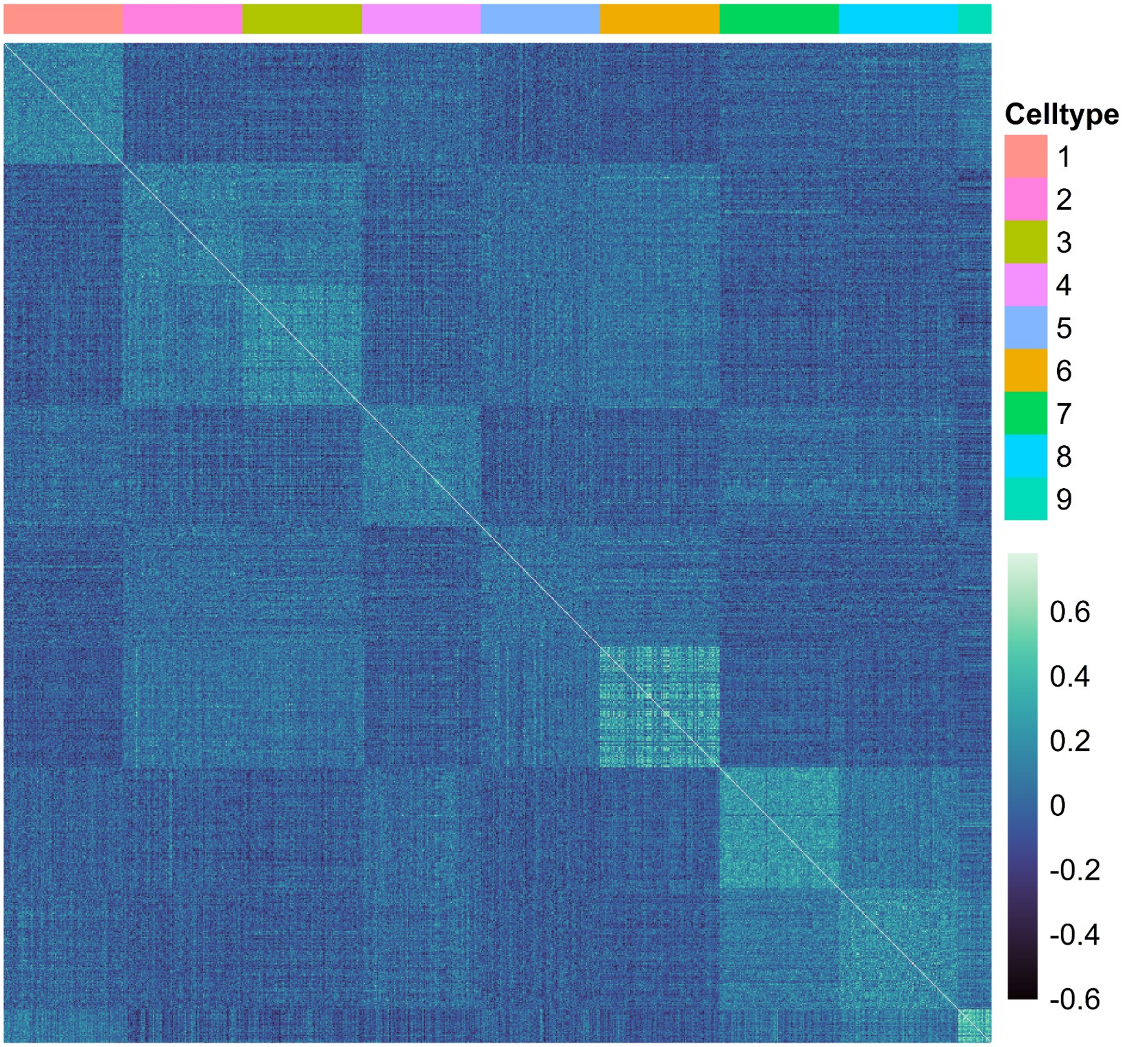


Figure S7: Cell-cell correlations in Dataset 4. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

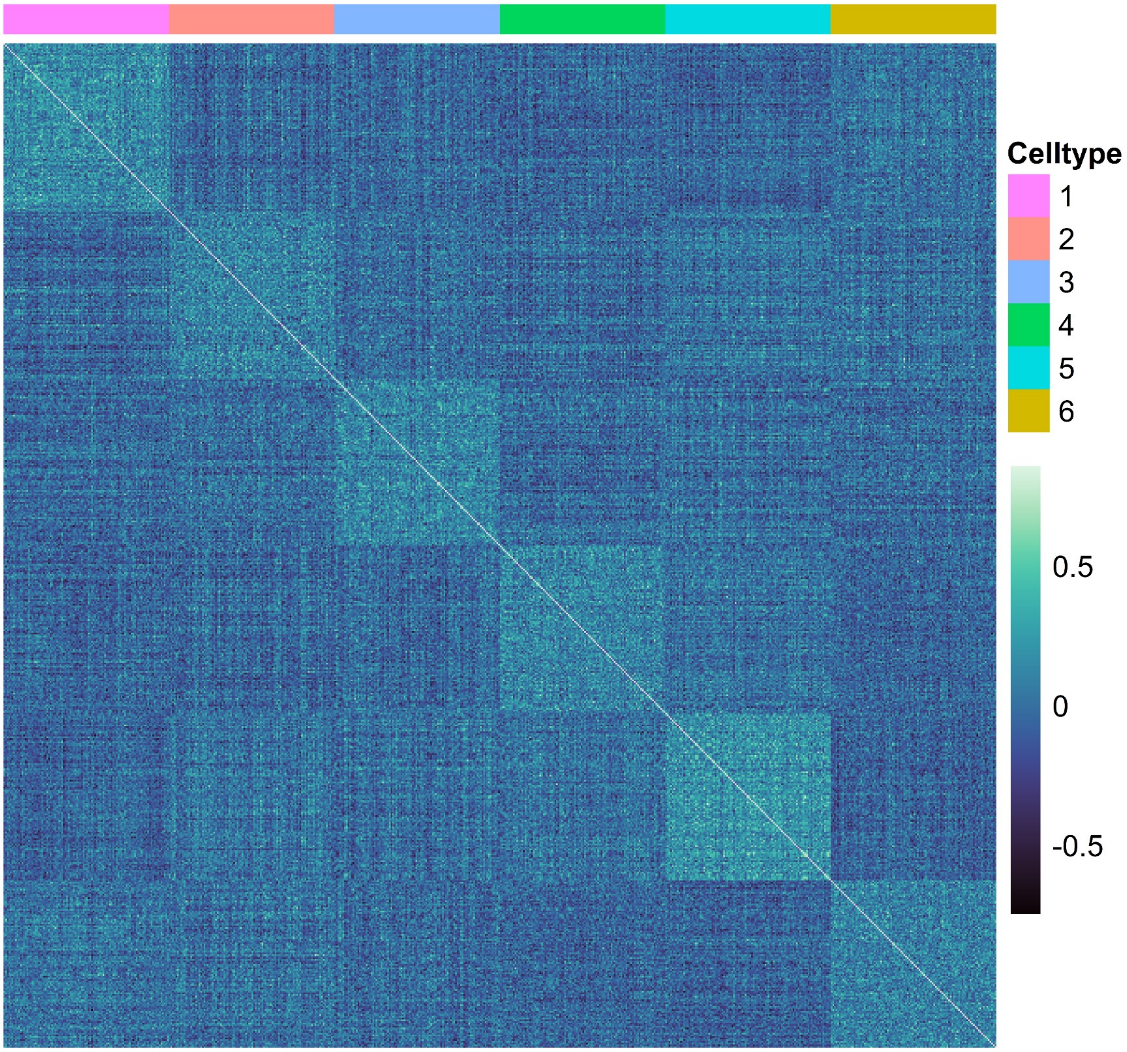


Figure S8: Cell-cell correlations in Dataset 5. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

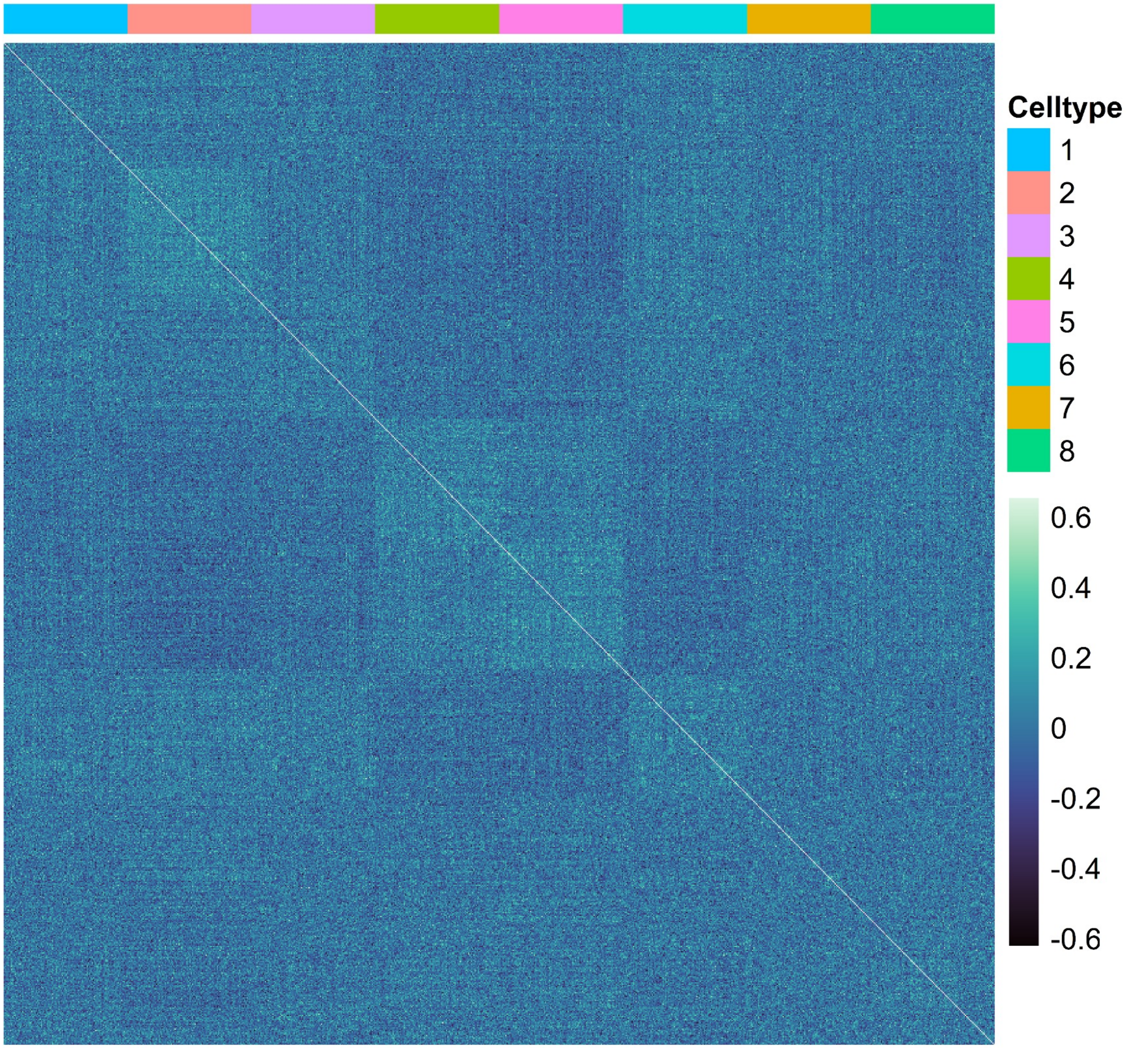


Figure S9: Cell-cell correlations in Dataset 6. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

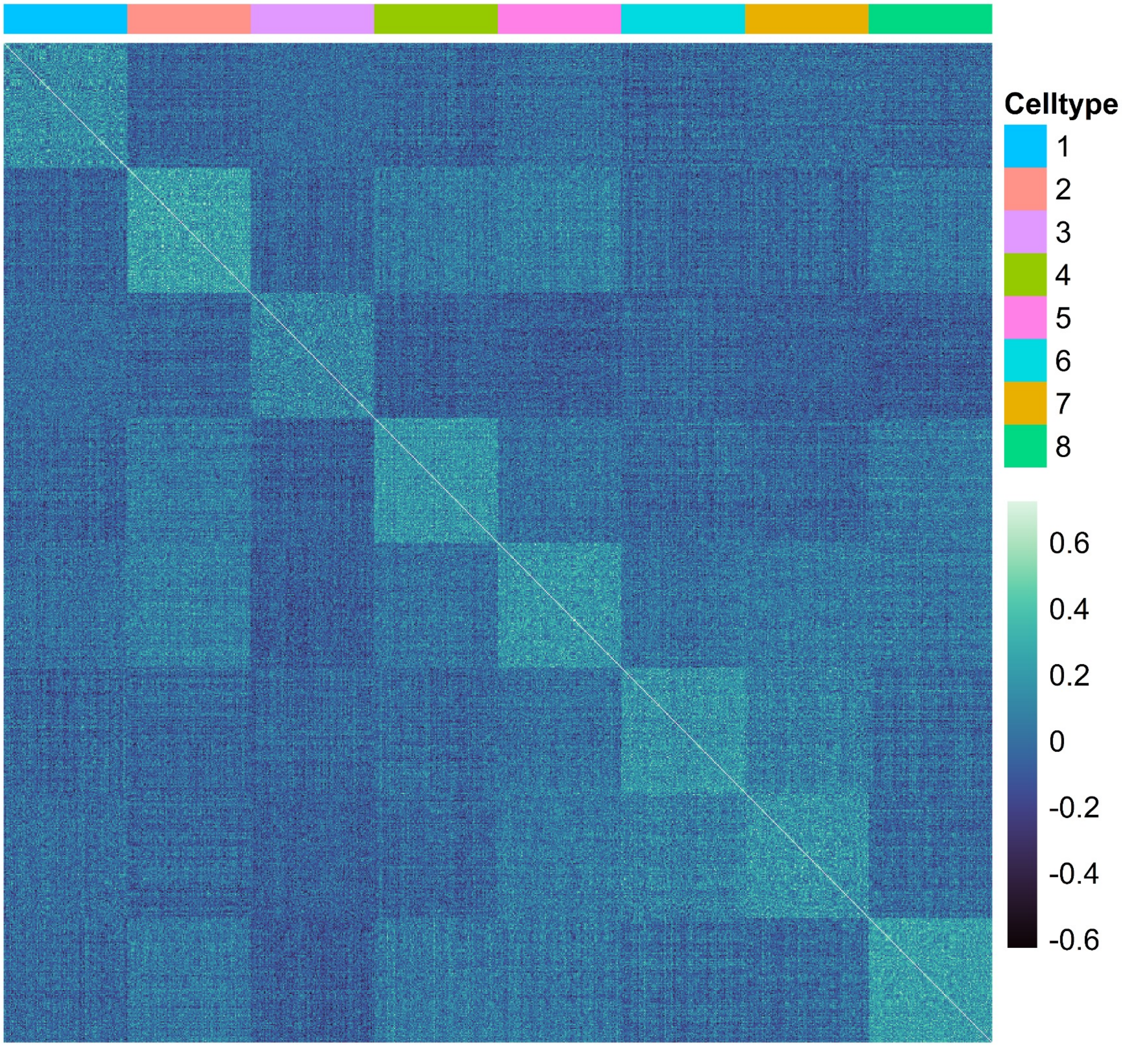


Figure S10: Cell-cell correlations in Dataset 7. Correlations are calculated based on the scaled gene expression levels of top 2000 highly variables identified by Seurat. For cell types with more than 100 cells, we randomly selected 100 cells to reduce the size of the heatmap.

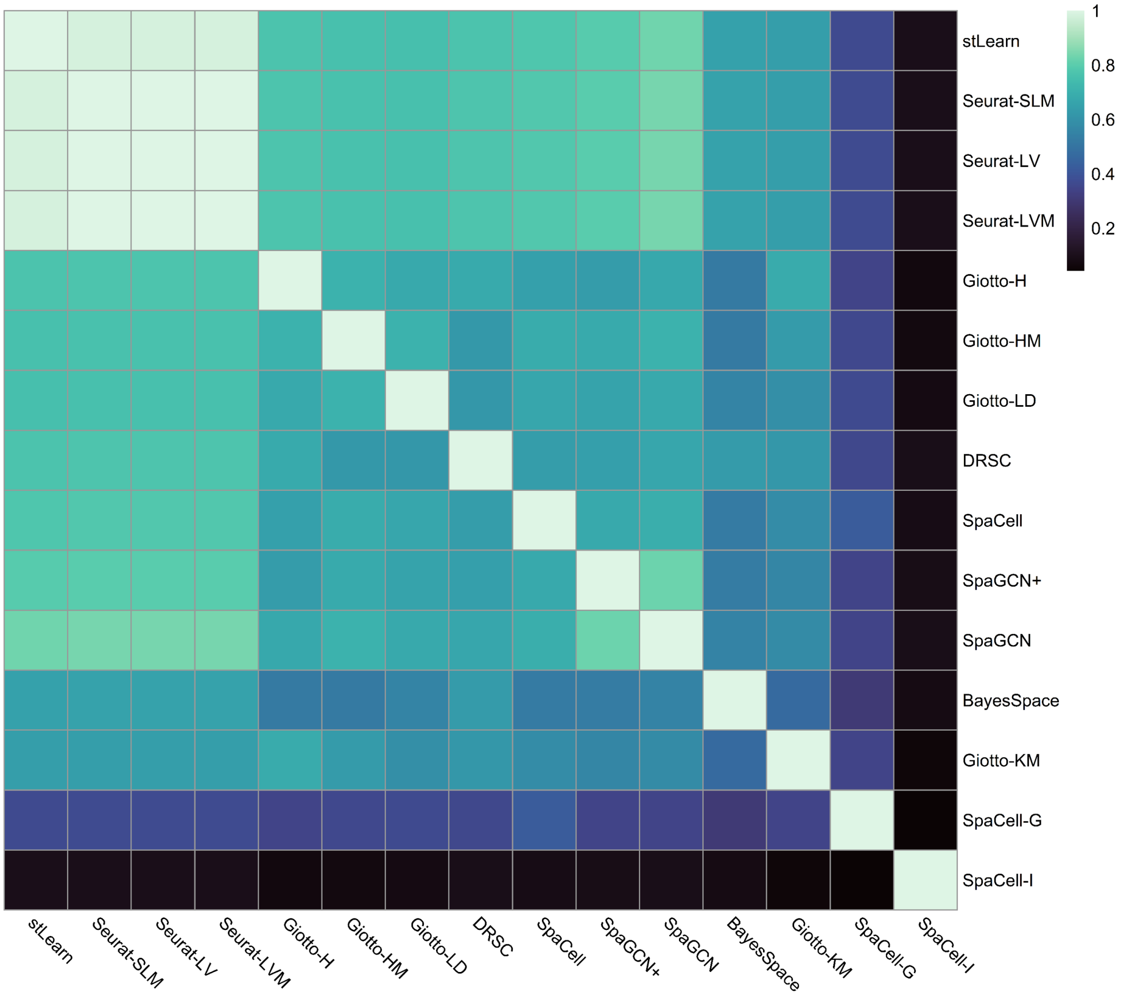


Figure S11: Concordance between different methods on Dataset 1. Concordance was measured by the ARI score between two sets of inferred cluster labels.

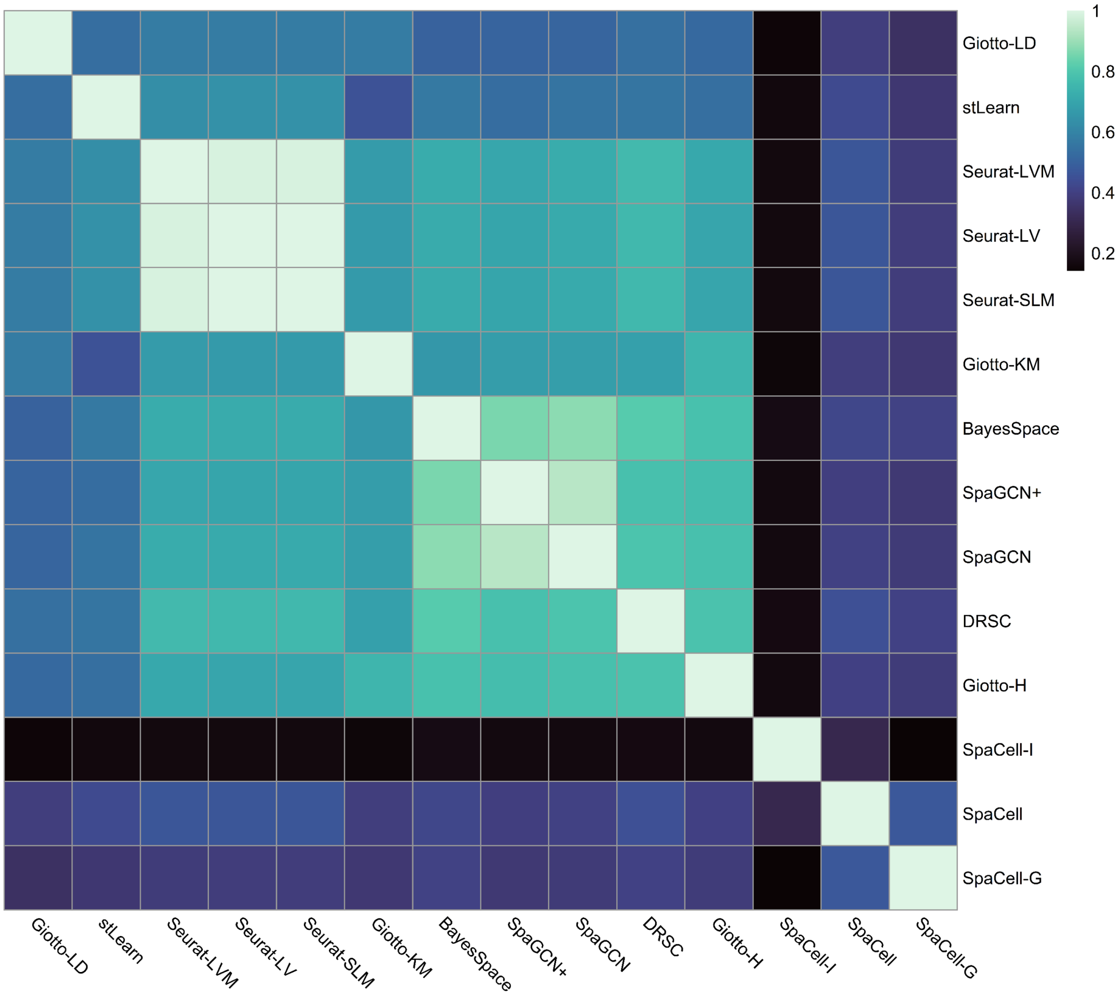


Figure S12: Concordance between different methods on Dataset 2. Concordance was measured by the ARI score between two sets of inferred cluster labels.

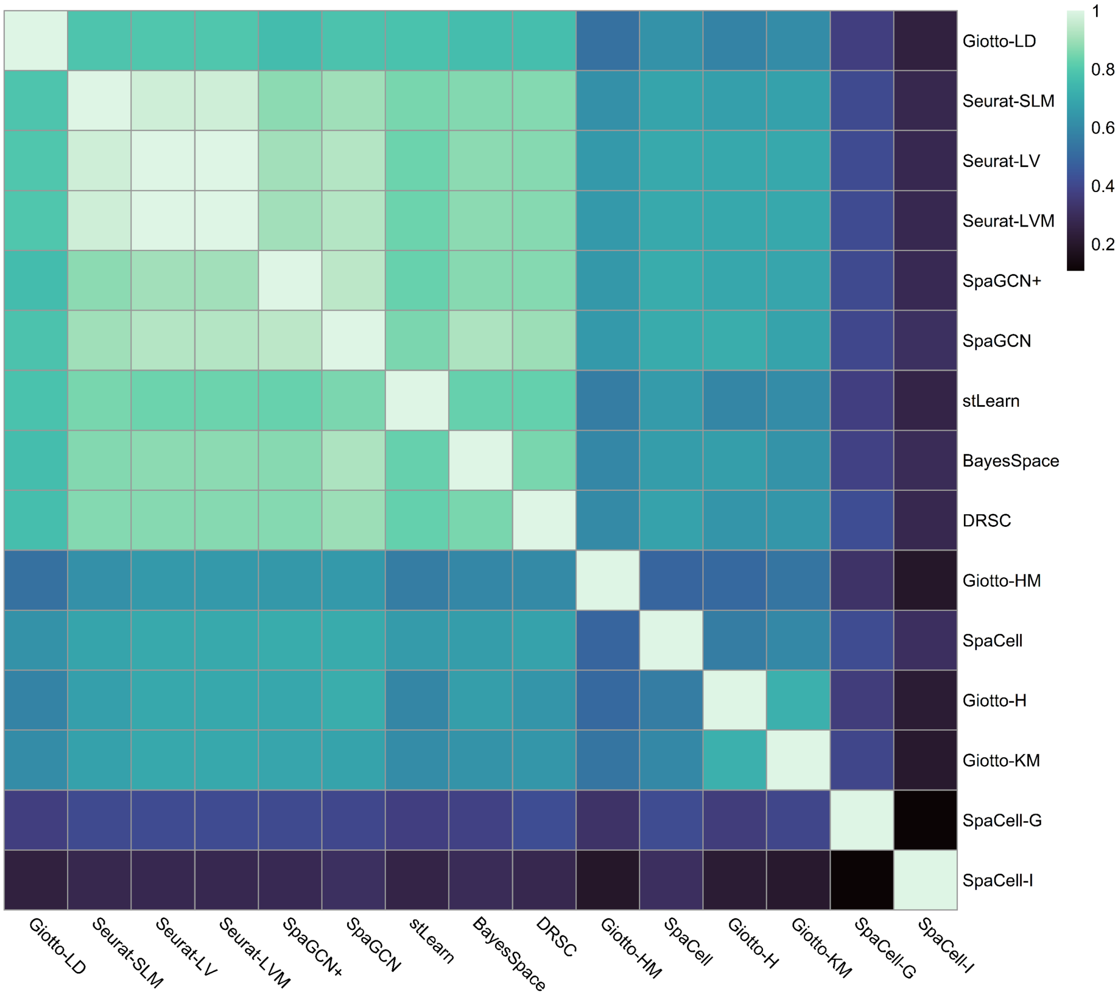


Figure S13: Concordance between different methods on Dataset 3. Concordance was measured by the ARI score between two sets of inferred cluster labels.

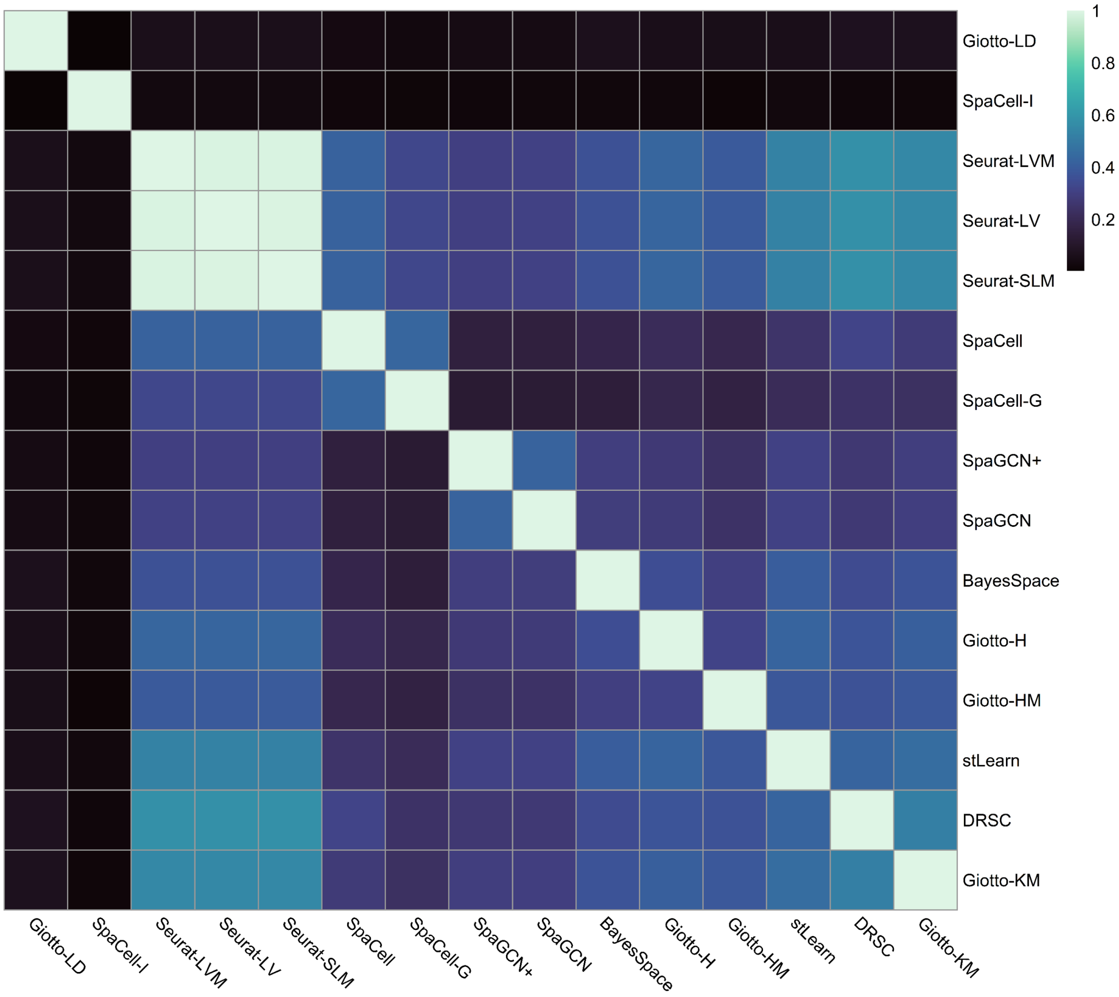


Figure S14: Concordance between different methods on Dataset 4. Concordance was measured by the ARI score between two sets of inferred cluster labels.

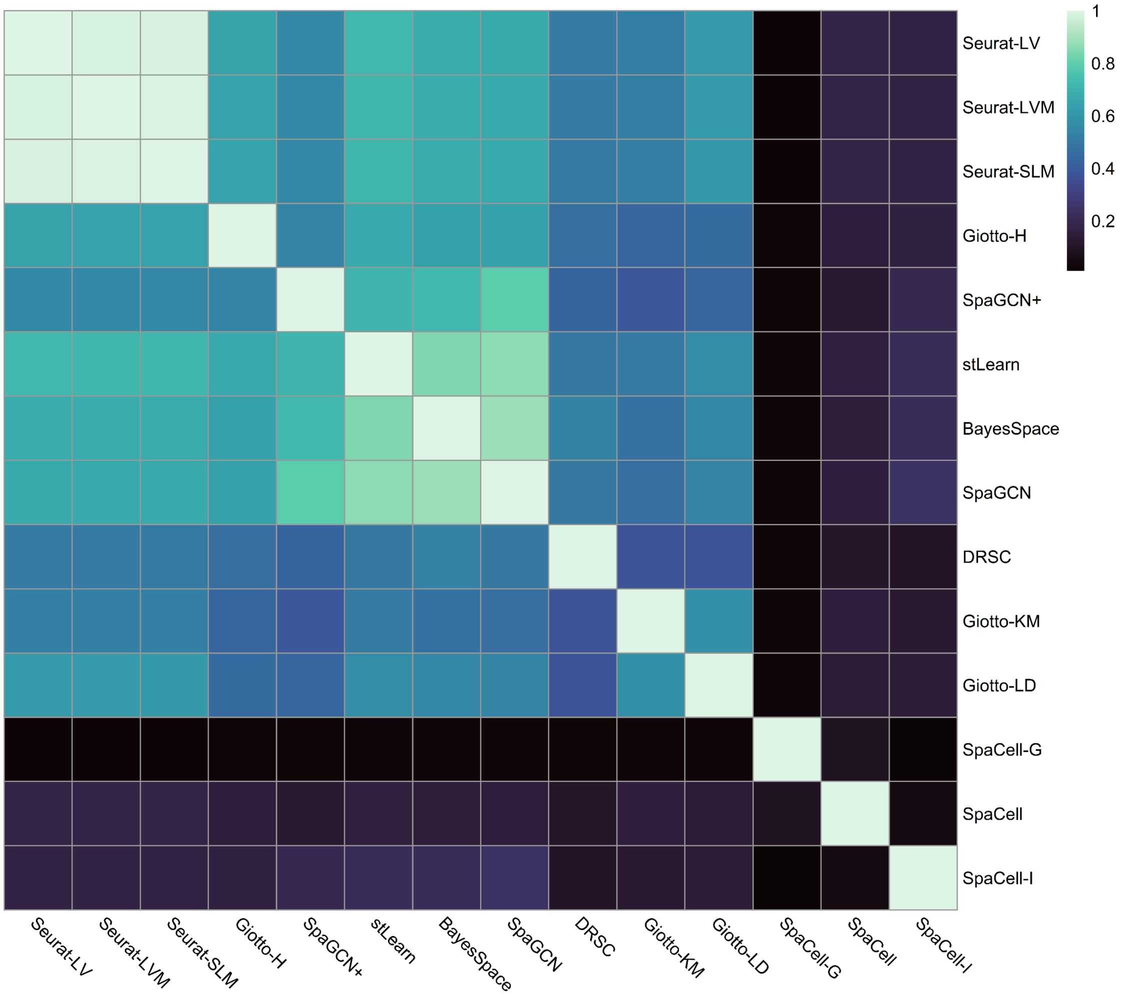


Figure S15: Concordance between different methods on Dataset 5. Concordance was measured by the ARI score between two sets of inferred cluster labels.

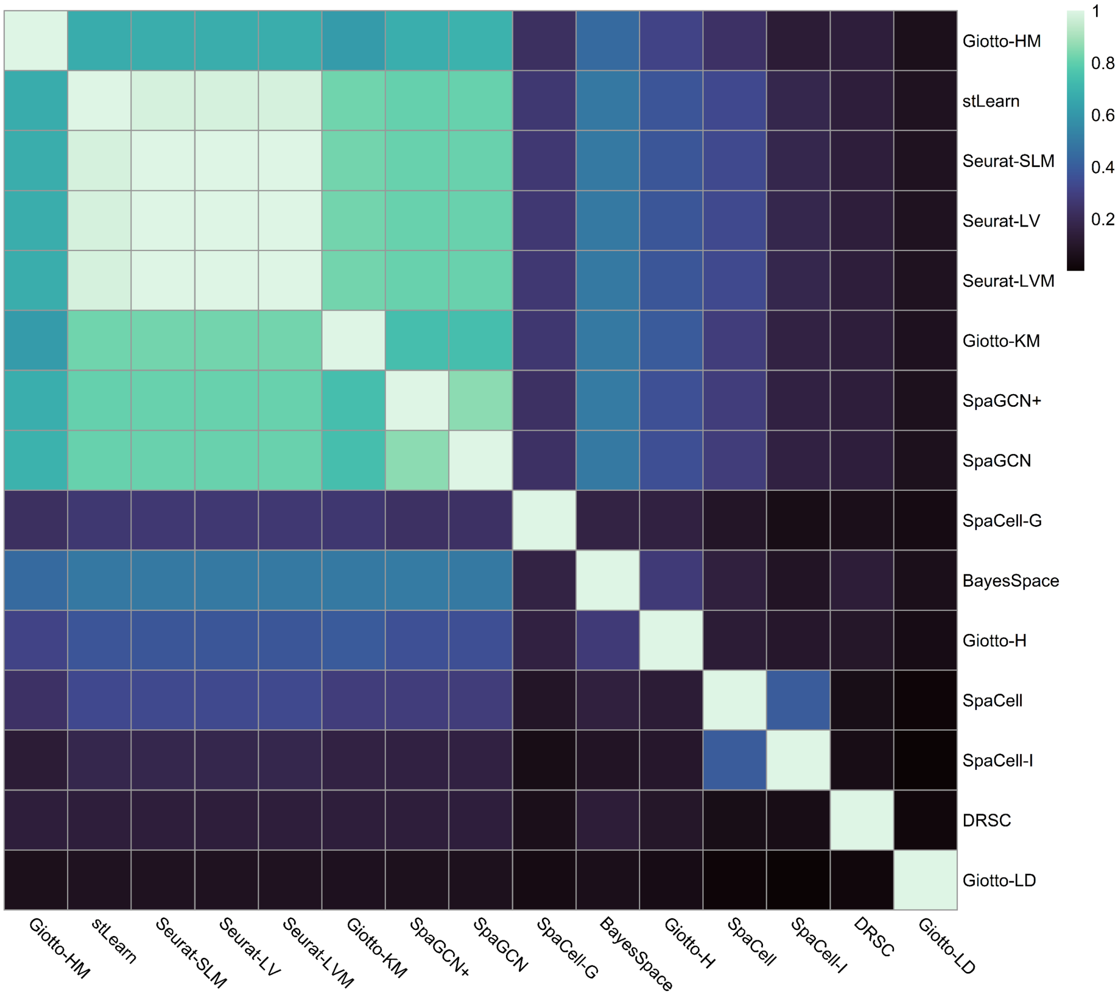


Figure S16: Concordance between different methods on Dataset 6. Concordance was measured by the ARI score between two sets of inferred cluster labels.

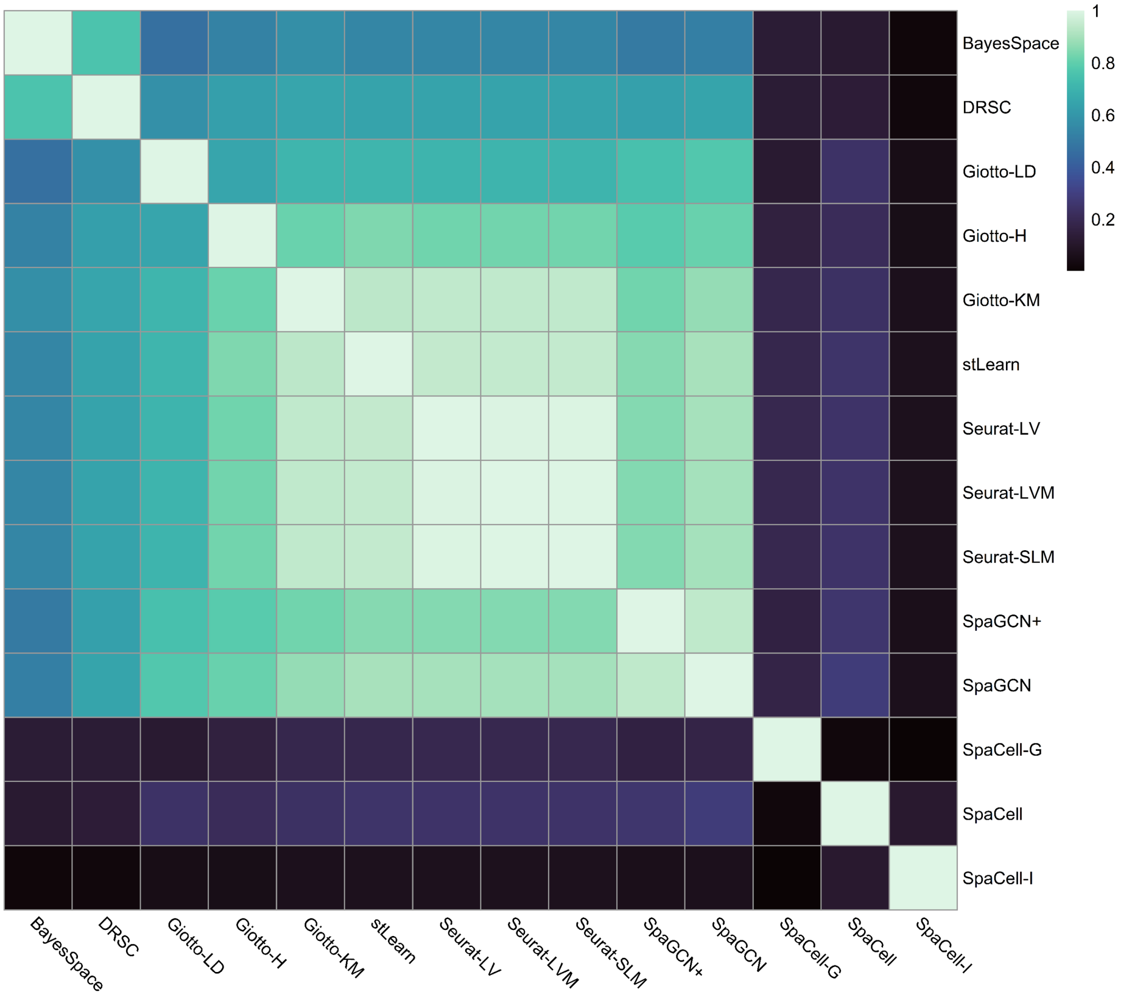


Figure S17: Concordance between different methods on Dataset 7. Concordance was measured by the ARI score between two sets of inferred cluster labels.

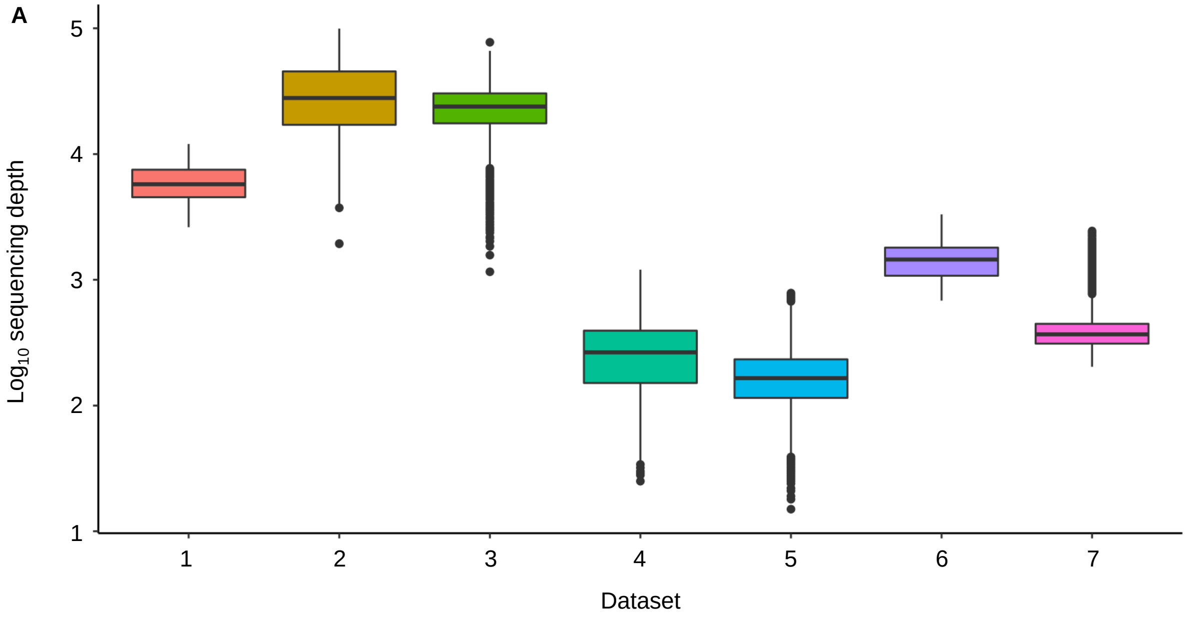


Figure S18: Log$_{10}$ sequencing depths of the first replicate in each dataset.

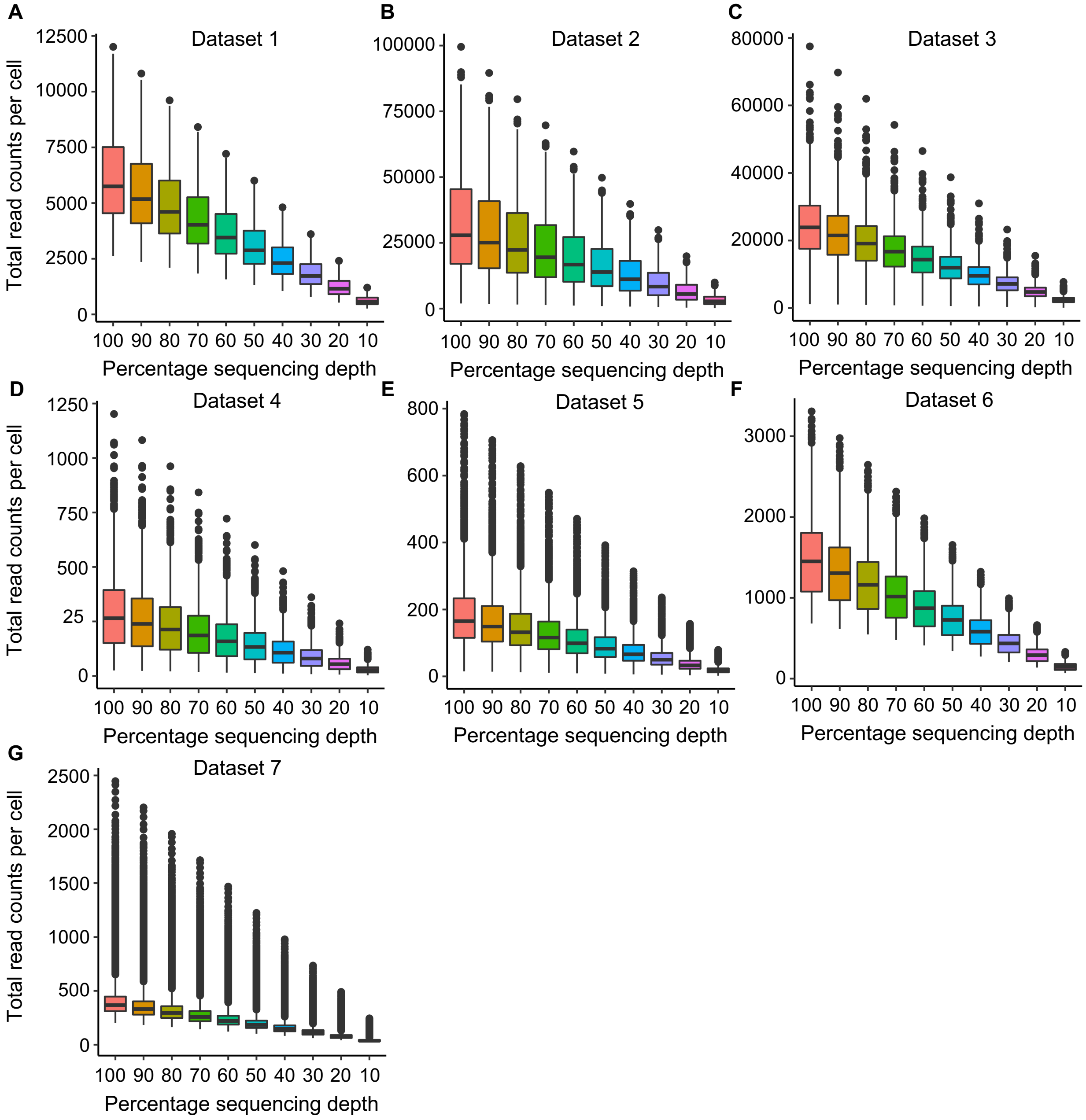


Figure S19: Comparison of sequencing depths after downsampling.

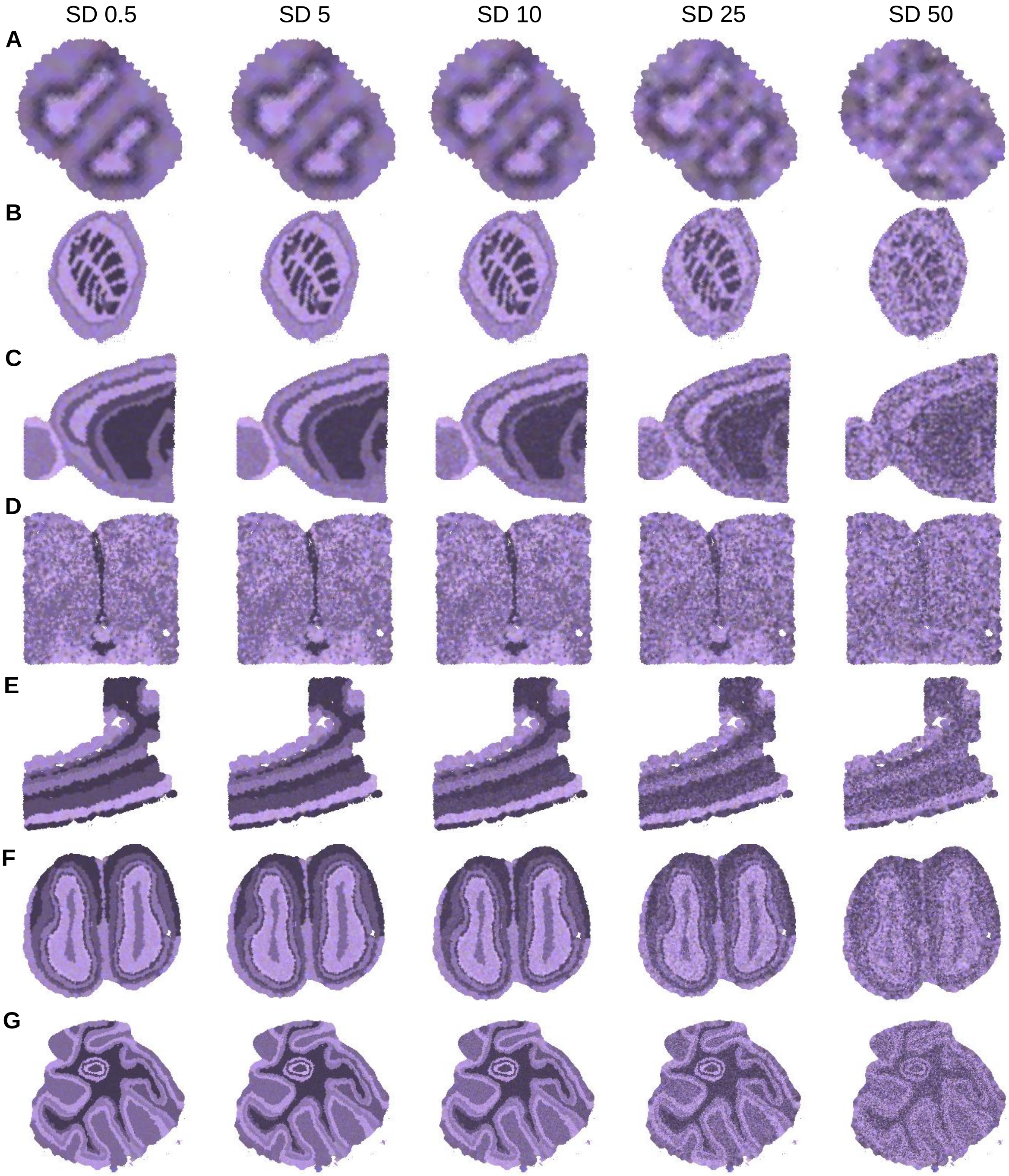


Figure S20: Simulated histology images with increasing standard deviations of pixel colors, ranging from 0.5 to 50. (**A-G**): Simulated images corresponding to Datasets 1-7, respectively.

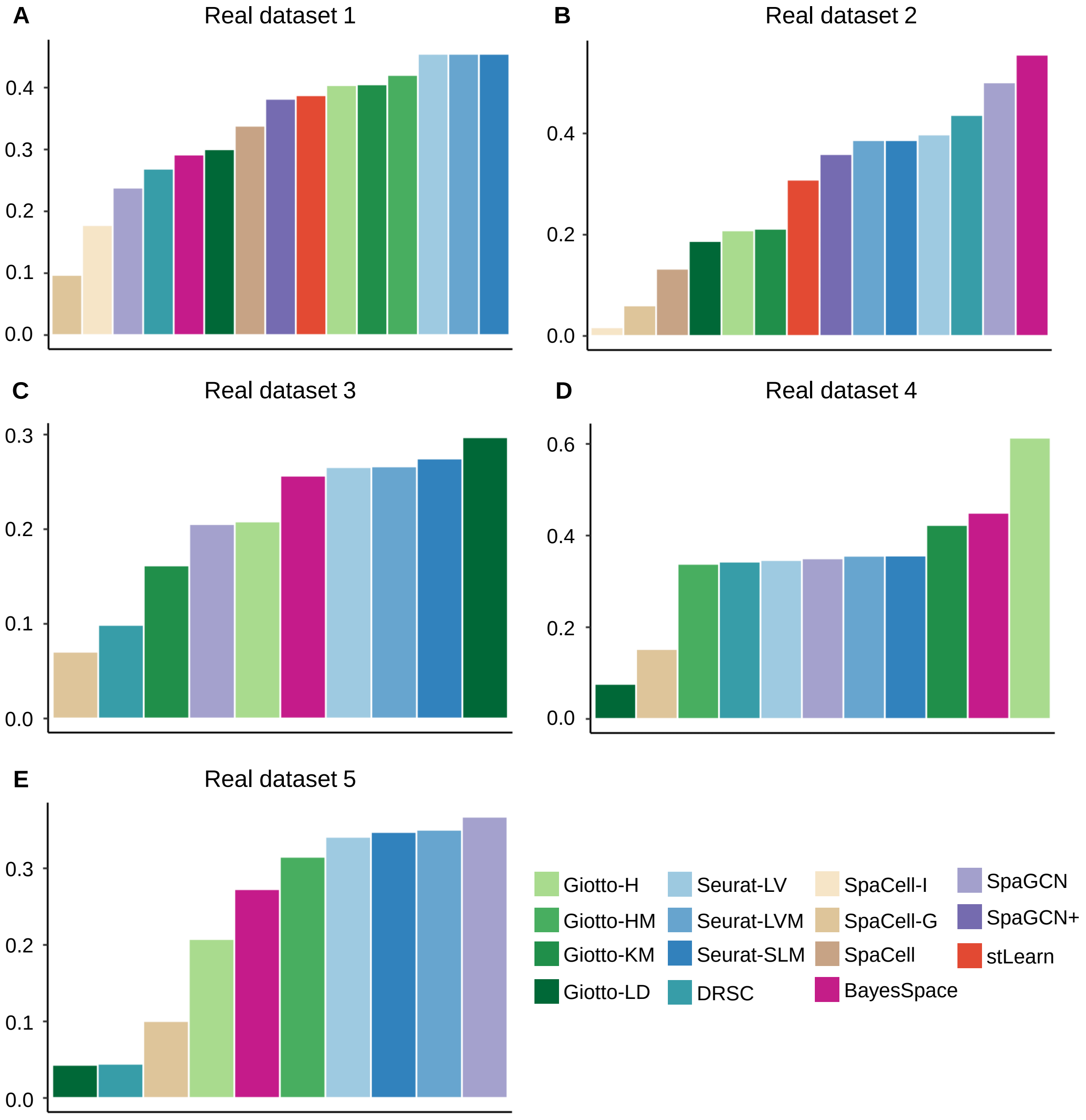


Figure S21: Comparison of clustering accuracy based on five real spatial transcriptomics datasets. (**A-E**): Adjusted Rand index (ARI) scores for real datasets 1-5.

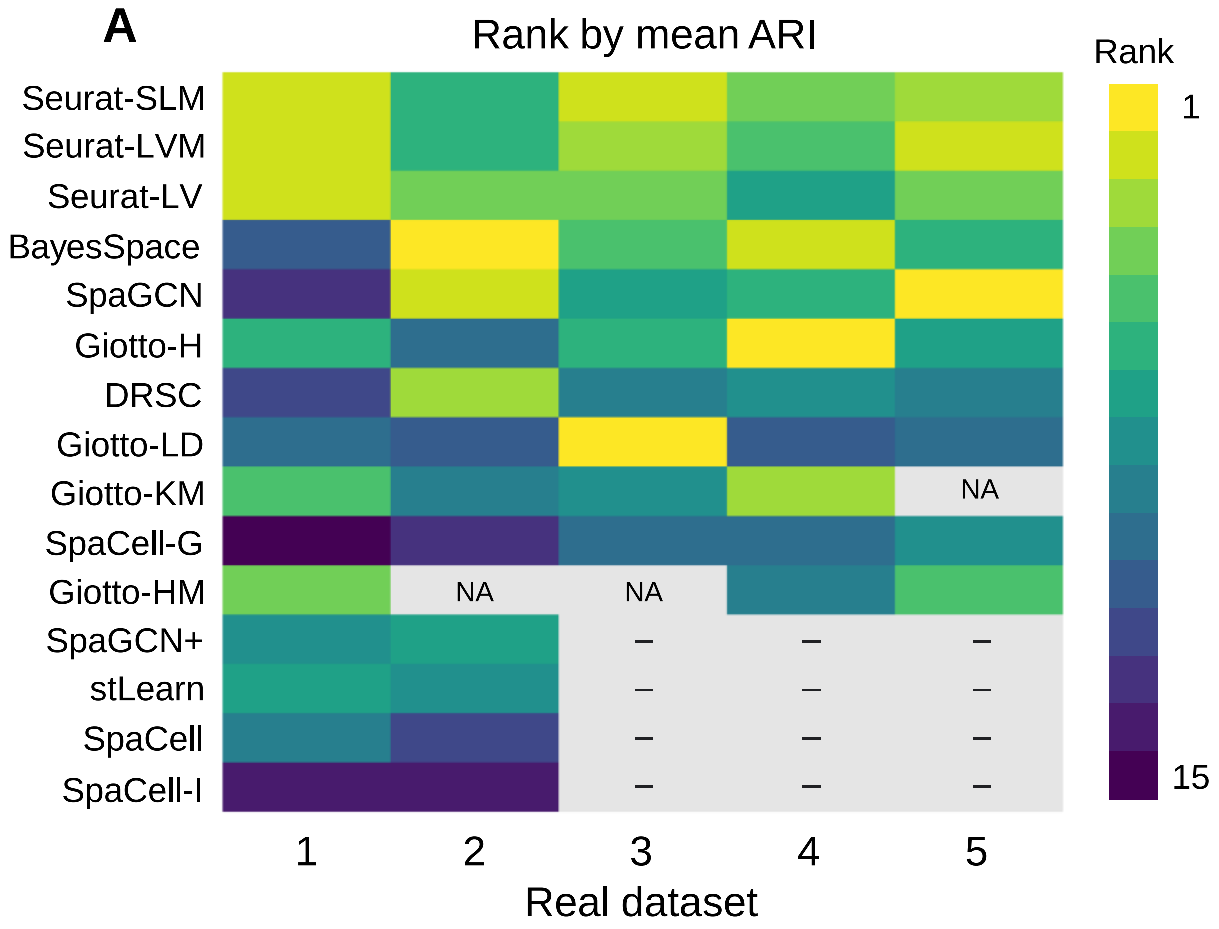


Figure S22: Ranking of methods based on ARI scores across the five real spatial transcriptomics datasets. Entries marked by NA indicate that the method encountered an error. Methods that require histology images as the input are not applicable to datasets 3-5 since no images are available, and the corresponding entries are marked by “-”.

Supplementary Table

Table S1: Summary of the real data characteristics.

Dataset	Technology	# of cells	# of genes	# of true cell types
Real dataset 1	Spatial Transcriptomics	265	16573	5
Real dataset 2	10X Genomics Visium	3611	33538	7
Real dataset 3	osmFISH	4839	33	12
Real dataset 4	MERFISH	5488	160	9
Real dataset 5	Stereo-seq	10000	26145	10

References

[1] Vladimir Y Kiselev, Kristina Kirschner, Michael T Schaub, et al. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14:483–486, 2017.

[2] Monika Krzak, Yordan Raykov, Alexis Boukouvalas, Luisa Cutillo, and Claudia Angelini. Benchmark and parameter sensitivity analysis of single-cell rna sequencing clustering methods. *Frontiers in genetics*, page 1253, 2019.

[3] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[4] Jeffrey R Moffitt, Dhananjay Bambah-Mukku, Stephen W Eichhorn, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.

[5] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, pages 1–10, 2021.

[6] Simone Codeluppi, Lars E Borm, Amit Zeisel, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nature Methods*, 15:932–935, 2018.

[7] Ao Chen, Sha Liao, Mengnan Cheng, Kailong Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jin Yang, Jiangshan Xu, Shijie Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.