**Methods**

*Simulated scRNA datasets by SPARSim*

To further evaluate the performances of the clustering methods, we generated simulated datasets using a state-of-the-art simulation algorithm SPARSim (Baruzzo *et al.*, 2020). Based on the previous benchmark study of simulation methods for single-cell RNA sequencing data (Cao *et al.*, 2021), SPARSim showed strong applicability for simulating multiple cell groups, which meets the requirement of our clustering task. Meanwhile, it outperforms other simulation methods on several data property evaluations, such as producing similar library size, mean expression, and cell correlation to the original template, preserving both technical and biological zeros, *etc*.

During the simulation, SPARSim first estimates the gene expression level intensities (), gene expression level variabilities (), and library size () from a given template to simulate a dataset with similar properties. Then, the expression level () of gene  in cell  is modeled using a Gamma distribution as

At the same time, the algorithm simulates the technical variability () for the expression of cell to represent the systematic and sample-dependent bias, which is modeled using a Multivariate Hypergeometric distribution as

In this study, we used the Darmanis dataset from the previous analyses and the Zheng dataset as the labeled templates (Zheng et al., 2017). The parameters estimated from the Zheng dataset are preset in the algorithm, and we estimated the above three parameters (*i.e.*, intensities, variabilities, and library size) for the Darmanis dataset using its raw data. The simulated datasets are summarized in **Table 1**. To evaluate the GMM-VGAE clustering stability of different sizes’ datasets, we simulated additional datasets with 1694 cells and 6776 cells using the Zheng preset by halving and doubling the cell numbers of each cell type, respectively. As for the simulated genes, we selected the top 1200 genes from the original templates via VST as same as before to make the simulated datasets consistent with the real datasets. Finally, we generated ten copies of datasets for each simulation and calculated the mean performances afterward.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Template | Darmanis | Zheng | Zheng | Zheng |
| Parameters preset in SPARSim? | No | Yes | Yes | Yes |
| Number of labeled clusters | 8 | 4 | 4 | 4 |
| Original cell numbers in template | 420 | 3388 | 3388 | 3388 |
| Simulated cell numbers | 420 | 3388 | 1694 | 6776 |
| Original gene numbers in template | 22085 | 15041 | 15041 | 15041 |
| Simulated gene numbers | 1200 | 1200 | 1200 | 1200 |
| Number of copies | 10 | 10 | 10 | 10 |

**Table 1**. The summary of simulated datasets

**References**

Cao, Y., Yang, P., & Yang, J. (2021). A benchmark study of simulation methods for single-cell RNA sequencing data. *Nature communications*, *12*(1), 6911.

Baruzzo, G., Patuzzi, I., & Di Camillo, B. (2020). SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics (Oxford, England)*, *36*(5), 1468–1475.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, *8*, 14049.