Quantitative Biology 2022, 10(3): 299–300
https://doi.org/10.15302/J-QB-022-0304

COMMENTARY

Integrating human single-cell data from multiple sources

Chenwei Li, Zedao Liu, Zemin Zhang*

BIOPIC, Peking University, Beijing 100871, China
* Correspondence: zemin@pku.edu.cn

Received June 2, 2022; Revised June 5, 2022; Accepted June 6, 2022

Single-cell RNA sequencing (scRNA-seq) has been a powerful tool for biomedical research and the number of scRNA-seq datasets has been growing rapidly thanks to the continuous advancement of library preparation technologies. In addition to the increasing number of cells being profiled, there is a trend of conducting cross-tissue analyses which build on the initial efforts that studied one tissue at a time. The Human Cell Atlas [1] and Human BioMolecular Atlas Program [2] aim to systematically characterize the expression profiles of various human organs and cell types, and to form comprehensive references for single-cell transcriptome data. However, due to the different data sources, how to effectively and consistently integrate these single-cell transcriptome datasets have remained as a great challenge. Specifically, there are three issues that need to be addressed. First, traditional relational databases cannot meet the requirements of efficient storage and retrieval of scRNA-seq dataset. Second, novel indexing methods are required so that each single cell could be traced with multiple attributes. Finally, standardized controlled vocabulary is required to annotate cell types in a unified manner.

Recently, Dr. Xuegong Zhang's group from Tsinghua University in China introduced hECA (human Ensemble Cell Atlas) [3] which has made important contributions to addressing the above issues. hECA is a human integrated cell atlas covering 38 organs and 11 systems, including transcriptomic data of more than 1 million cells from over 100 single-cell research studies. Based on a unified information framework, hECA enables seamless cell-centric data integration from different sources. In terms of database architecture, hECA used a unified giant table (uGT) as the storage engine. It is worth noting that uGT can not only cover transcriptome features such as gene expression, but also store various types of cell attributes, thereby supporting cell-level information retrieval based on multiple attributes. To further make cell types from different data sources comparable, they also developed a unified hierarchical annotation framework (uHAF) as the underlying indexing system. In addition to providing controlled vocabulary with hierarchy and affiliation for cell type annotation, uHAF is compatible with the widely used Cell Ontology database and supports continuous updates in the future. In order to facilitate users to use the retrieval function in the system, the authors also developed the ECAUGT Python package for programmatic access. Based on these modules, the authors proposed three application scenarios and demonstrate their reliability. First, users can select cells from the virtual body through flexible logical expressions, such as filtering cells of interest based on cell type names or expression thresholds of marker genes. Second, users can quantitatively view the relationship among gene expression levels, cell types, and organs through its visualization module. The user-friendly data visualization website will be very helpful for researchers with non-biological information backgrounds. Finally, users can customize reference creation for automatic cell type classification.

In conclusion, hECA has made a landmark contribution to integrating human single-cell data from multiple sources and performing downstream analysis. In addition to the convenience for the majority of scientific researchers to use the integrated single-cell

data for analysis, the innovations in its data structure and controlled vocabulary for cell type annotation represent a wealth for the bioinformatics community. As an evolving project, we believe that future versions of hECA will integrate more datasets and provide more capabilities for data analysis.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Chenwei Li, Zedao Liu and Zemin Zhang declare that they have no conflict of interest or financial conflicts to disclose.

OPEN ACCESS

REFERENCES

1. Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., *et al*. (2017) The human cell atlas. eLife, 6, e27041
2. Snyder, M. P., Lin, S., Posgai, A., Atkinson, M., Regev, A., Rood, J., Rozenblatt-Rosen, O., Gaffney, L., Hupalowska, A., Satija, R., *et al*. (2019) The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature, 574, 187−192
3. Chen, S., Luo, Y., Gao, H., Li, F., Chen, Y., Li, J., You, R., Hao, M., Bian, H., Xi, X., *et al*. (2022) hECA: The cell-centric assembly of a cell atlas. iScience, 25, 104318

