Journal Pre-proof

CDCP: a visualization and analyzing platform for single-cell datasets

Yuejiao Li, Tao Yang, Tingting Lai, Lijin You, Fan Yang, Jiaying Qiu, Lina Wang, Wensi Du, Cong Hua, Zhicheng Xu, Jia Cai, Zhiyong Li, Yiqun Liu, Ling Li, Minwen Zhang, Jing Chen, Lei Zhang, Dongsheng Chen, Weiwen Wang, Shiping Liu, Liang Wu, Wenjun Zeng, Bo Wang, Xiaofeng Wei, Longqi Liu, Fengzhen Chen

PII: S1673-8527(21)00374-X

DOI: https://doi.org/10.1016/j.jgg.2021.12.004

Reference: JGG 999

To appear in: *Journal of Genetics and Genomics*

Received Date: 1 September 2021

Revised Date: 9 December 2021

Accepted Date: 12 December 2021

Please cite this article as: Li, Y., Yang, T., Lai, T., You, L., Yang, F., Qiu, J., Wang, L., Du, W., Hua, C., Xu, Z., Cai, J., Li, Z., Liu, Y., Li, L., Zhang, M., Chen, J., Zhang, L., Chen, D., Wang, W., Liu, S., Wu, L., Zeng, W., Wang, B., Wei, X., Liu, L., Chen, F., CDCP: a visualization and analyzing platform for single-cell datasets, *Journal of Genetics and Genomics*, https://doi.org/10.1016/j.jgg.2021.12.004.

CDCP: a visualization and analyzing platform for single-cell datasets

Recently, the rapid advances of single-cell sequencing technologies, including sequencing of single-cell genomics, transcriptomes, epigenetics, and spatial transcriptomes, have empowered researchers to scrutinize cellular heterogeneity, gene expression, epigenetic modifications and spatial information of gene expression at single-cell resolution, which also leads to a continuous accumulation of massive single-cell datasets (Clark et al., 2016; Gawad et al., 2016). However, the scarcity of a suitable platform for analyzing single-cell datasets is a major impediment for in-depth mining of the existing data.

Currently, there are few databases available that integrate only different aspects of datasets (Taverna et al., 2020; Longo et al., 2021; Zhao et al., 2021). For example, CancerSEA, a database covering single-cell transcriptomics of 41,900 cells from 25 cancer types, aims to explore the different functional states of cancer cells at a single-cell resolution (Yuan et al., 2019). Abugessaisa et al. (2018) established SCPortalen, a database of publicly available human and mouse single-cell datasets. Consequently, regardless of these databases, there is still an immense need to build a comprehensive platform encompassing single-cell datasets covering different cell types of all species. Likewise, the continuous gathering of single-cell datasets has brought challenges for the research community, especially for bench scientists, to process and even conduct routine analyses of their own massive datasets. However, learning bioinformatics would definitely help to navigate and process these complex datasets, but it could be very time-consuming and costly. Therefore, it is imperative to establish a user-friendly dataset analysis platform that will meet the individual needs of bench researchers.

Here, we develop a freely accessible (https://db.cngb.org/cdcp/) Cell-omics Data Coordinate Platform (CDCP) to share and integrate single-cell datasets. It allows visualization of each single-cell dataset with tSNE cell dimensionality reduction graphs, cluster analysis graphs of different cell types, and histograms displaying the number of

different cell types. In addition, the expression patterns of multiple genes in different cell types or clusters can be displayed with cluster and violin graphs. Moreover, CDCP contains detailed information on the biological samples of the datasets and allows downloading the raw sequences and expression matrices of each single-cell dataset. CDCP, unlike other platforms such as the UCSC single cell browser and ASAP, provides a platform to share and integrate single-cell transcriptomics datasets and allows users to upload their data to make it a real-time updated platform. Importantly, personalized analysis is also supported in the CDCP.

CDCP is established and maintained using the Django web framework (https://www.djangoproject.com/), the Python programming language and Pycharm (https://www.jetbrains.com/pycharm/). Centos7 (https://www.centos.org/) is chosen as the operating system of the server. In terms of system architecture, we use NGINX (http://nginx.org/) to provide static resource access, uWSGI (https://uwsgi-docs.readthedocs.io/en/latest/) to deploy query and download services, and Postgresql to store metadata. In terms of data security, the database is deployed in the CNGB (Wang et al., 2019), which has passed the three-level review of information security level protection and the protection capability review of trusted cloud services. Moreover, all services have been deployed with high availability. A user-friendly analysis pipeline is developed in the platform to facilitate the reanalysis of single-cell datasets of interest. By providing a single-cell expression matrix, users can calculate quality control metrics, annotate highly variable genes, perform principal component analysis, compute a neighborhood graph of observations, embed the neighborhood graph and find marker genes for characterizing groups.

Database content and statistics. CDCP aims to provide a platform for sharing single-cell datasets and allows query of the expression of different genes in specific cell types or clusters. By now CDCP contains expression profiles of 6467 samples and 474,573 cells from public datasets in species including humans, monkeys and other animals. Although the existing datasets are collected from only 21 studies, the application of this

platform will attract a large number of researchers to upload and share their datasets on it.

User interface. A series of user-friendly interfaces are used to support various functions of CDCP. It provides five main functions: data exploration, associated databases, bioinformatics analysis, data visualization and data submission (Fig. 1A). On the "home" page, users can quickly have a view of the statistical data of the projects, samples, organisms and cells covered by CDCP (Fig. 1B). By clicking the buttons of "Project", "Sample", "Organism" or "Cell", users can be directed to the "Explore" page for data retrieval. A summary of datasets for diverse organisms is presented below. Users can obtain the number of projects, samples, cells, and organs provided in different categories (Fig. S1A). In addition, a brief introduction of shared databases, including MBA, NHPCA, HCL and VThunter, is provided for users (Fig. S1A).

Data exploration. On the "Explore" page, one can retrieve a group of cells by selecting the attributes of the Project, Organism, Tissue, Disease, Library strategy or Release date. The query results that meet the search criteria will be displayed, including detailed information about the datasets (Fig. S1B). Actions for downloading raw data and visualizing results are also available. For data not available for download, users can apply to the submitter and download it with permission. CDCP supports visualization for data submitted with an expression matrix. If only the original sequencing data are submitted, the platform does not support visualization now. When clicking a specific project ID, users can be directed to the corresponding page that displays the detailed meta information of each single-cell sample. This includes the following information: the project name summarizing the type of the dataset, the abstract of the relevant publication, the submitter and affiliation of the dataset, and the metadata for downloading as a text file. Moreover, statistical information, including sample size, experiments and runs, is presented. By clicking on the number of samples, users can access detailed information about each sample, including age, gender, and origin. Similarly, click the number of experiments to obtain the detailed information of each

experiment, such as library strategy and library layout.

Associated single-cell databases. CDCP also provides external links to databases of MBA, NHPCA, HCL and VThunter in the "Databases" module. The Macaque Brain Atlas (MBA) depicts the macaque brain atlas and allows searching for gene expression in diverse brain cell types at single-cell resolution. The Non-Human Primate Cell Atlas (NHPCA) is a single-cell transcriptomics database that provides visualization and analysis of transcriptomics and epigenetic single-cell datasets sampled from non-human primate organs or tissues. Human Cell Landscape (HCL) illustrates a basic scheme of the human cell landscape to identify the cell type composition of human organs using single-cell RNA-seq. VThunter allows single-cell screening of virus receptor expression in the whole animal kingdom and predicts the host tropism of various viruses. These available databases will provide more comprehensive information for users interested in brain research, human, non-human primate models, and virus research.

Data analysis. On the "Analysis" page, the description and workflow of Codeplot, the analyzing tool, is introduced (Fig. S1C). To facilitate the analysis and sharing of single-cell transcriptome sequencing datasets, a single-cell workspace was implemented to manage relevant datasets and perform comprehensive bioinformatics analysis and visualization based on the expression matrix datasets archived in CDCP. By clicking the "single-cell workspace", users can be directed to the Codeplot page, which provides a computing platform with a trusted execution environment for bioinformatics analysis. Users only need to provide a single cell expression matrix (csv/tsv) for data analysis. Sufficient optional adjustment parameters are provided to support user-defined parameters. Each step will generate the hdf5 file as input for the next step. We also provide each step as a workflow for users to debug.

Data visualization. Visualization of the selected dataset using tSNE analysis is displayed on the "Visualization" page, with dot colors representing different cell

clusters or cell types (Fig. 1C). Users can browse the cell clusters and cell numbers of each cell type in different projects by selecting a project and a dataset in the corresponding drop-down box presented in the hierarchical navigation menu (Fig. 1D and 1E). Moreover, by selecting the gene of interest in the "Primary Gene List" drop-down box, users can analyze the expression pattern of this gene in all clusters and cell types (Figs. 1F, 1G, S1D). In addition, the EXHIBIT Double-positive function was established to describe the co-expression patterns of two genes in different cell types and clusters with the exhibition of a tSNE plot and violin plot (Fig. S1E), which can be used to reveal functional gene annotation and transcriptional regulatory programs.

Data submission. CDCP allows scientists to submit their own single-cell transcriptomics datasets for sharing with the research community. Single-cell data submission supports five main entities: project, sample, experiment, run, and analysis data. The submission of single-cell datasets is mainly divided into two processes: submission of original sequencing data and submission of analysis results. By clicking "The submission of single-cell data" on the "Submission" page, users will be directed to the submission portal page (Fig. S1F). Users need to fill detailed information of each sample. Before submitting data files, project and sample information should be submitted. The detailed steps for submitting datasets and the data type and format for submission are available on the "Submission" page to guide users. For large files, it is recommended to submit data through aspera, which supports resumable upload.

Example application. To demonstrate the potential application of CDCP, we take the query of the human breast immune cell dataset as an example. By selecting "homo sapiens" in the "Organism" drop-down box and "breast" in the "Tissue" drop-down box, a project that met the search criteria was presented (Fig. S1B). This project depicts a single-cell atlas of 9683 immune cells from 14 triple-negative breast cancer samples (Qiu et al., 2019). By clicking the "Visualization" button on the right, users can be directed to the Visualization page. As shown by the tSNE plot, 22 cell clusters were obtained and annotated as T cells (9 clusters), macrophages (6 clusters), B cells (3

clusters), dendritic cells (2 clusters) and 2 unassigned clusters. T cells are the dominant group of immune cells, followed by macrophages, B cells and dendritic cells. The different characteristics of T cell clusters can be revealed by known markers and differentially expressed genes (DEGs) among clusters. For instance, among the nine clusters of T cells, clusters T7 and T9 exhibited features of regulatory T cells with high expression of *FOXP3* (ENSG00000049768) and *IL2RA* (ENSG00000134460) (Figs. S1D and S2A). The DEGs were identified among clusters, such as the expression of CD200 (ENSG00000091972) in cluster T8, the expression of *SOX4* (ENSG00000124766) in cluster T9 and the expression of *ZNF683* (ENSG00000176083) in T6 (Fig. S2B-S2D). These results indicate that CDCP is a user-friendly, reliable and practical platform for in-depth mining of existing single-cell datasets.

CDCP aims to provide users with a comprehensive resource of published single-cell datasets to make them easily accessible to biology researchers. As of July 2021, CDCP has collected single-cell databases from 6,467 samples and 474,573 cells. This enables retrieval and visualization of single-cell datasets and the expression profiles of a specific gene in different cell clusters or cell types. In addition, CDCP also allows visualization of the co-expression patterns of two genes for scientists interested in studying gene correlations. CDCP contains single-cell RNA-seq and ATAC-seq of different cells from humans, monkeys and other animals. This may help scientists in diverse research fields retrieve datasets of interest in a more convenient way. In addition to providing a comprehensive resource for single-cell datasets, an additional strength of CDCP is the supply of the network analysis toolkit (CodePlot), which facilitates customized data processing by users. Additional work will focus on integrating more data sources. Our future goal is to update the database with more single-cell datasets to cover more species, including animals, plants, and microbes. Despite the advantages, there are some limitations of the CDCP database; for example, not all single-cell datasets are available for download. This depends on whether the submitter has enabled downloadable permissions. Additionally, only a few datasets support visualization because some of the submitted datasets lack the expression matrix necessary for

automatic visualization. Further improvements will be made to ensure the submission of the expression matrix.

In summary, CDCP provides a substantial source for the research community to access and analyze single-cell datasets of interest. The diverse functions of CDCP, such as visualization and analyzing tools, may help biologists attain valuable information from the available public as well as their own datasets.

Conflict of interest

The authors declare no competing interests.

Acknowledgments

This work was supported by the China National GeneBank (CNGB).

References

Clark, S.J., Lee, H.J., Smallwood, S.A., Kelsey, G.,Reik, W., 2016. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol. 17, 72.

Gawad, C., Koh, W.,Quake, S.R., 2016. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 17, 175-188.

Longo, S.K., Guo, M.G., Ji, A.L.,Khavari, P.A., 2021. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat Rev Genet. 22, 627-644.

Qiu, S., Hong, R., Zhuang, Z., Li, Y., Zhu, L., Lin, X., Zheng, Q., Liu, S., Zhang, K., Huang, M., Lee, K., Lu, Q., Xia, W., Xu, F., Wang, X., Tang, J., Xiao, X., Wei, W., Yuan, Z., Shi, Y., Hou, Y., Zhang, X., Wang, J., Yang, H., Zhan, Q., Li, B.,Wang, S., 2019. A Single-Cell Immune Atlas of Triple Negative Breast Cancer Reveals Novel Immune Cell Subsets. bioRxiv. 566968.

Taverna, F., Goveia, J., Karakach, T.K., Khan, S., Rohlenova, K., Treps, L., Subramanian, A., Schoonjans, L., Dewerchin, M., Eelen, G.,Carmeliet, P., 2020. BIOMEX: an interactive workflow for (single cell) omics data interpretation and visualization. Nucleic Acids Res. 48, W385-W394.

Wang, B., Liu, F., Zhang, E.C., Wo, C.L., Chen, J., Qian, P.Y., Lu, H.R., Zeng, W.J., Chen, T., Wei, J.P., Wan, Q., Wang, R.,Xu, X., 2019. [The China National GeneBank horizontal line owned by all, completed by all and shared by all]. Yi Chuan. 41, 761-772.

Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H.,

Long, Z., Shi, A., Zhao, T., Xiao, Y.,Li, X., 2019. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res. 47, D900-D908.

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., Hao, J.,Peng, J., 2021. SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res. 49, D1413-D1419.

Yuejiao Li[1],

BGI-Shenzhen, Shenzhen 518083, China

Tao Yang[1], Tingting Lai[1], Lijin You[1], Fan Yang[1],

China National GeneBank, Shenzhen 518120, China

Jiaying Qiu[1],

BGI-Shenzhen, Shenzhen 518083, China

College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Lina Wang, Wensi Du, Cong Hua, Zhicheng Xu, Jia Cai, Zhiyong Li, Yiqun Liu, Ling Li, Minwen Zhang, Jing Chen, Lei Zhang,

China National GeneBank, Shenzhen 518120, China

Dongsheng Chen,

BGI-Shenzhen, Shenzhen 518083, China

College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

Weiwen Wang,

China National GeneBank, Shenzhen 518120, China

Shiping Liu, Liang Wu,
BGI-Shenzhen, Shenzhen 518083, China

Wenjun Zeng, Bo Wang, Xiaofeng Wei[*]*,*
China National GeneBank, Shenzhen 518120, China

Longqi Liu[*]*,*
BGI-Shenzhen, Shenzhen 518083, China

Fengzhen Chen[*]
China National GeneBank, Shenzhen 518120, China

[1] *These authors contributed equally*
***** *Corresponding authors.*
Email addresses: weixiaofeng@cngb.org (X. Wei), liulongqi@genomics.cn (L. Liu), chenfengzhen@cngb.org (F. Chen)

Figure legend

Fig. 1. Overview and application of CDCP. **A**: Overview of the CDCP platform. **B**: The homepage of CDCP illustrating the covered single-cell datasets and five main functions. **C**: Visualization of the dataset with a tSNE cell dimensionality reduction graph. **D**: The distribution of cell clusters in different cell types. **E**: Cell numbers in different clusters are displayed. **F**: tSNE graph showing the expression profile of *FOXP3* (ENSG00000049768) in different clusters. **G**: The expression of *FOXP3* (ENSG00000049768) in different cell types.