SCIENCE CHINA
Life Sciences

•RESEARCH PAPER•

https://doi.org/10.1007/s11427-022-2224-4

Privacy-preserving integration of multiple institutional data for single-cell type identification with scPrivacy

Shaoqi Chen[1†], Bin Duan[1†], Chenyu Zhu[1], Chen Tang[1], Shuguang Wang[1], Yicheng Gao[1], Shaliu Fu[1], Lixin Fan[4*], Qiang Yang[4*] & Qi Liu[1,2,3*]

[1] *Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Orthopaedic Department of Tongji Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China;*
[2] *Translational Medical Center for Stem Cell Therapy and Institution for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China;*
[3] *Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai 201210, China;*
[4] *Department of AI, WeBank, Shenzhen 518055, China*

Received August 14, 2022; accepted September 15, 2022; published online December 15, 2022

The rapid accumulation of large-scale single-cell RNA-seq datasets from multiple institutions presents remarkable opportunities for automatically cell annotations through integrative analyses. However, the privacy issue has existed but being ignored, since we are limited to access and utilize all the reference datasets distributed in different institutions globally due to the prohibited data transmission across institutions by data regulation laws. To this end, we present *scPrivacy*, which is the first and generalized automatically single-cell type identification prototype to facilitate single cell annotations in a data privacy-preserving collaboration manner. We evaluated *scPrivacy* on a comprehensive set of publicly available benchmark datasets for single-cell type identification to stimulate the scenario that the reference datasets are rapidly generated and distributed in multiple institutions, while they are prohibited to be integrated directly or exposed to each other due to the data privacy regulations, demonstrating its effectiveness, time efficiency and robustness for privacy-preserving integration of multiple institutional datasets in single cell annotations.

Citation: Chen, S., Duan, B., Zhu, C., Tang, C., Wang, S., Gao, Y., Fu, S., Fan, L., Yang, Q., and Liu, Q. (2022). Privacy-preserving integration of multiple institutional data for single-cell type identification with scPrivacy. Sci China Life Sci 65, https://doi.org/10.1007/s11427-022-2224-4

INTRODUCTION

Single-cell transcriptomics is indispensable for understanding cellular mechanisms of complex tissues and organisms (Guan et al., 2021; Jiang et al., 2021; Plass et al., 2018; Xie et al., 2021; Zhao et al., 2022). As single-cell technologies rapidly developed over recent years, its experimental throughput increased substantially, allowing to profile increasingly complex and diverse samples, and accumulating vast numbers of datasets over time. Integrative analyses of such large-scale datasets originating from various samples, different platforms and different institutions globally, offer unprecedented opportunities to establish a comprehensive picture of cell landscape. To this end, various community generated large-scale atlas-level single cell reference data, such as the Human Cell Atlas (HCA) (Regev et al., 2017), Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020), BRAIN Initiative Cell Census Network (Winnubst and Arber, 2021) , Human Lung Atlas (Travaglini et al., 2020), Human Gut Atlas (Elmentaite et al., 2020), Hu-

†Contributed equally to this work
*Corresponding authors (Qi Liu, email: qiliu@tongji.edu.cn; Qiang Yang, email: qyang@cse.ust.hk; Lixin Fan, email: lixinfan@webank.com)

man BioMolecular Atlas Program (HuBMAP) (Snyder et al., 2019), The Tabula Sapiens (Jones et al., 2022), hECA (Chen et al., 2022a) etc., and recently great achievements has been made in the building of pan-tissue single-cell transcriptome atlases covering more than a million cells, including 500 cell types, across more than 30 human tissues from 68 donors (Domínguez Conde et al., 2022; Eraslan et al., 2022; Jones et al., 2022; Liu and Zhang, 2022; Suo et al., 2022). These references data facilitate the automatically cell type annotations in a supervised way without prior marker gene annotations (Aran et al., 2019; Duan et al., 2020; Kiselev et al., 2018; Li et al., 2020; Liu et al., 2020; Ma and Pellegrini, 2020; Stuart et al., 2019). It is obviously that integrating more reference datasets or combining these atlas-level data will improve the cell type annotations (Aran et al., 2019; Duan et al., 2021; Kiselev et al., 2018; Stuart et al., 2019), and various integration methods for single cell annotations have been presented (Aran et al., 2019; Chen et al., 2022b; Duan et al., 2020; Kiselev et al., 2018). However, all these existing integration methods require to access the relevant reference datasets directly, which may be unavailable due to the data privacy and security issues. Currently, the privacy and political issues towards omics data transmission and sharing across different institutions or countries are gradually attracting attentions. On the one hand, countries around the world are strengthening laws to protect data privacy and security by prohibition of certain data transition across countries or organizations. Such regulations include the General Data Protection Regulation (GDPR) (Politou et al., 2018) implemented by the European Union, the Health Insurance Portability and Accountability Act (HIPPA) (Benefield et al., 2006) and Health Information Technology for Economic and Clinical Health Act (HITECH) (Halamka and Tripathi, 2017) enacted by U.S. and etc. On the other hand, single-cell reference datasets are generated and accumulated rapidly in different institutions around the world. It is a great demand to integrate all these institutional data globally to facilitate the establishment of the comprehensive picture of human cell reference map. However, these institutions may be required to protect data privacy and security and prohibit certain data transmission across organizations and countries by data regulation laws (Table 1). As a result, there exist an inevitable contradiction between the rapidly accumulated single cell reference data and the privacy issue among data sharing and integration. Current references integrating strategies failed to address these issues, hindered by legal restrictions on data sharing (Lotfollahi et al., 2022). Moreover, current integrating methods ignore the problem of privacy towards single cell sequencing data, as scRNA-seq datasets of human are sensitive and likely to contain sufficient sequencing depth to call genetic variants (Byrd et al., 2020).

To this end, taking the single-cell transcriptome data as an initial study, we propose *scPrivacy*, an efficient, flexible and extendable automatically single-cell type identification prototype and a proof-of-concept study to facilitate single cell annotations in a data privacy-preserving collaboration manner, by integrating multiple references single cell transcriptome data distributed in different institutions using a federated learning based deep metric learning framework. Federated learning is a collaborative paradigm in privacy-preserving computing community that enables the institutions collaboratively to train a model while keeping the data in local institutions (Chen et al., 2021; Zhang and Yang, 2018). We summarized existing privacy-preserving methods with their distint characteristics and explained the necessity to perform privacy-preserving computing for large scale single cell data using federated learning framework (Table 2). In addition, in our previous studies (Duan et al., 2020; Duan et al., 2021), metric learning was also proven to be effective for single-cell type annotation. Briefly, the basic idea of *scPrivacy* is to make each institution train their models locally and aggregate encrypted models parameters for all institutions to avoid putting raw data of all institutions together directly. We evaluated *scPrivacy* on a comprehensive set of 27 publicly available benchmark datasets for single cell type identification to stimulate the scenario that the reference datasets are rapidly generated and accumulated from multiple institutions, as well as on 15 publicly available patients datasets to simulate a large-scale real world situation that multiple hospitals collaborate together to build an automated cell type annotation system for COVID-19 patients, while they are prohibited to be integrated directly or exposed to each other due to the data privacy regulations, and demonstrated its effectiveness, time efficiency and robustness for privacy-preserving integration of multiple institutional datasets.

RESULTS

Overview of *scPrivacy*

scPrivacy is an efficient, flexible and extentable automatically single-cell type identification prototype to facilitate single cell annotations in a data privacy-perserving collaboration manner, by integrating multiple references single cell transcriptome data distributed in different institutions using a federated learning based deep metric learning framework. *scPrivacy* can effectively integrate information from multiple references while keeping each reference in local institutions, so as to solve the problem of data privacy protection. In particular, each institution trains its model on its local dataset and sends encrypted model parameters to server and then the server aggregates the parameters and sends back the aggregated model to institutions iteratively in training process. Specifically, *scPrivacy* comprises two main steps: model learning and cell assign-

Table 1 Single cell atlases and their corresponding data regulation laws

Atlas	Data regulation laws
Human Cell Atlas (Regev et al., 2017)	GDPR
Expression Atlas (Papatheodorou et al., 2019)	EMBL Internal Policy for Data Protection
Human Tumor Atlas Network (Rozenblatt-Rosen et al., 2020)	NIH Genomic Data Sharing Policy
BRAIN Initiative Cell Census Network (Winnubst and Arber, 2021)	NIH Genomic Data Sharing Policy
Human Lung Atlas (Travaglini et al., 2020)	All U.S. Federal, state and local laws and regulations
Human Gut Atlas (Elmentaite et al., 2020)	All U.S. Federal, state and local laws and regulations
HuBMAP (Hu, 2019)	HuBMAP External Data Sharing Policy
The Tabula Sapiens (Jones et al., 2022)	The Tabula Sapiens Privacy Policy including GDPR

Table 2 Current privacy-preserving methods and their properties

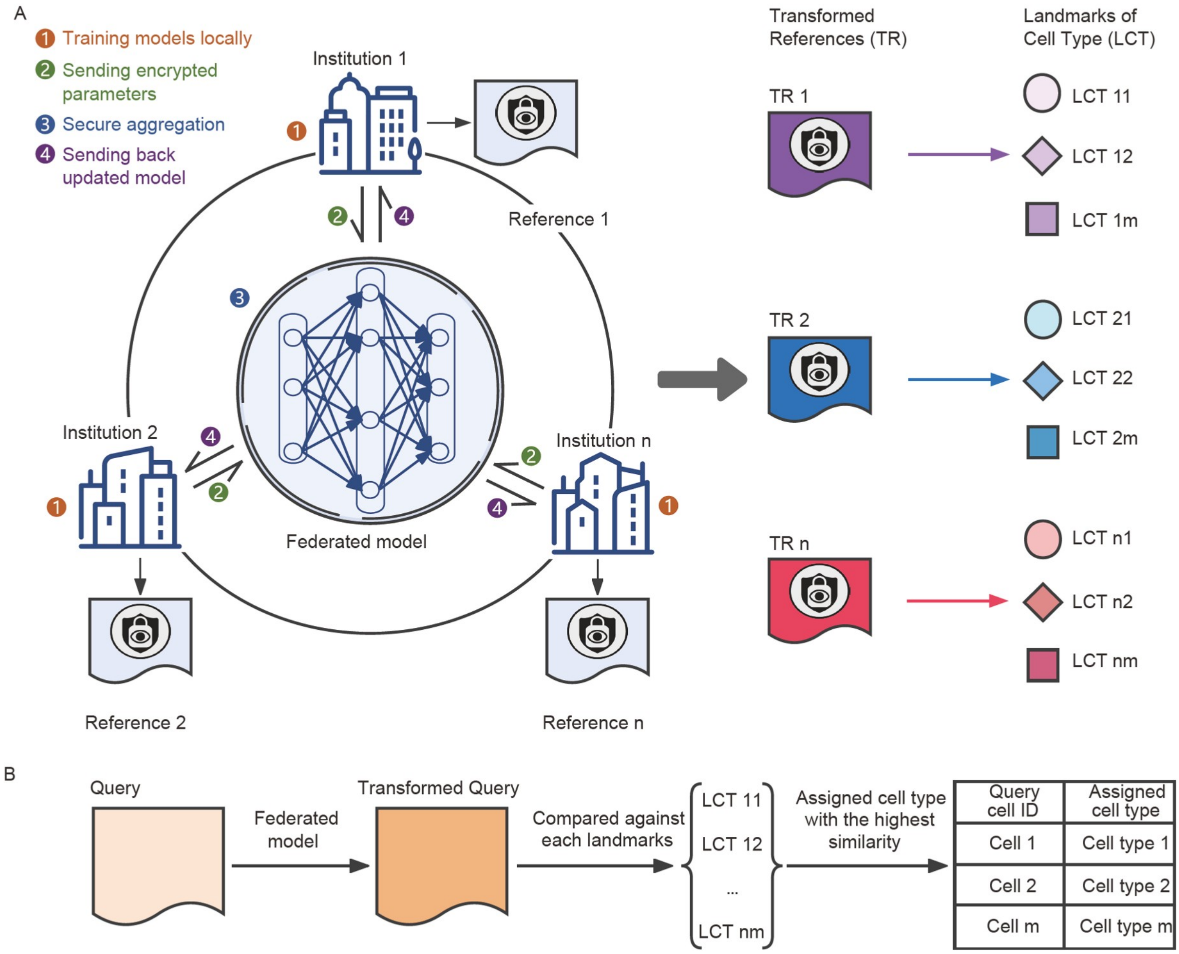
	Cleartext	Software Guard Extensions (McKeen et al., 2016)	Homomorphic Encryption (Acar et al., 2019)	Secure Multi-Party Computation (Yao, 1982)	Federated learning (Yang et al., 2019)
Risk for calling genetic variants?	Yes	No	No	No	No
Encrypted raw data?	No	No	Yes	Yes	No
Practicable for handling large scale single cell reference data?	Yes	No	No	No	Yes
Applicable to data transmission regulation laws?	No	No	No	No	Yes
Performance compared to cleartext based integration?	–	Identical	Identical	Identical	Upper bound

ment (Figure 1 and see Methods).

In the model learning stage (Figure 1A), *scPrivacy* trains a federated deep metric learning model on multiple institutional datasets in a data privacy-preserving manner. For an individual institution, deep metric learning (DML) is applied to learn an optimal measurement fitting the relationship among cells in the reference dataset, and the N-pair loss (Sohn, 2016) is used as the loss function for model training. With DML, cells belong to the same type became more similar and cells belong to different types became more dissimilar. Then, *scPrivacy* extends DML to a federated learning framework by aggregating model parameters of institutions to construct an aggregated model (Figure 1A and see Methods), which fully utilized the information contained in multiple institutional datasets to train the aggregated model while avoiding integrating datasets physically. In addition, *scPrivacy* can utilize the complementary information from different institutional datasets to boost the cell assignment performance, while also avoid the over correction of batch effect, as proven in previous study (Sohn, 2016). In the cell assignment stage (Figure 1B and see Methods), the query dataset is first transformed by the federated model to the same embedding space as that of the transformed institutional datasets. Then, the transformed query dataset is assigned to proper cell types by comparing with cell type landmarks of transformed institutional datasets. Specifically, for each transformed institutional dataset, *scPrivacy* carries out a cell search by measuring the similarity between the transformed query cells and cell type landmarks of the transformed institutional datasets. Finally, the query cells are assigned to the proper cell type with the highest similarity among all landmarks of the transformed institutional datasets.

Benchmarking *scPrivacy* with multiple institution and single institution

We firstly benchmarked *scPrivacy* with multiple institution (default), and compared it with that of *scPrivacy* with single institution to prove the benefits and necessity of integrating multiple institutional datasets. Here, "*scPrivacy* with single institution" represents that the results are achieved by training with one institution dataset and testing on another institution datasets. To this end, we collected 27 datasets (Table S1 in Supporting Information) from four studies of three tissues to simulate the data collaboration scenario among institutions: one study on the brain, one study on the pancreas and two studies on peripheral blood mononuclear cells (PBMCs) (Ding et al., 2019; Mereu et al., 2020). For brain tissue, the study contained four brain datasets (Tasic et al., 2016; Tasic et al., 2018) with different sources. For pancreas tissue, the study contained four commonly used pancreas datasets (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Xin et al., 2016). For PBMCs, the first study (Mereu et al., 2020) contained 12 datasets from 12 different sequencing platforms ("PBMC-Mereu"), and the



Query cell ID	Assigned cell type
Cell 1	Cell type 1
Cell 2	Cell type 2
Cell m	Cell type m

Figure 1 The *scPrivacy* workflow. A, The model learning process of *scPrivacy*. The federated model was trained with four steps: (1) training models locally; (2) sending encrypted parameters; (3) secure aggregation; (4) sending back updated model. Then cell type landmarks will be calculated for each transformed reference. B, The cell assignment process of *scPrivacy*. The federated model is utilized to transform the query cells. Then, the transformed query cells are compared against cell type landmarks of transformed institutional datasets, and the predicted cell type with the highest similarity among all cell type landmarks is obtained.

second study (Ding et al., 2019) contained 7 datasets from 7 different sequencing platforms (“PBMC-Ding”). In this case, each dataset of a tissue is simulated as an institution. For the strategy of *scPrivacy* with single institution, each dataset among multiple datasets was simulated as the dataset in an individual institution to train *scPrivacy* and the rest datasets were used as query datasets. For the strategy of *scPrivacy* with multiple institutions, each dataset among the multiple datasets was used as the query, and the others datasets were used to simulate distributed multiple institutional datasets and they are “virtually” integrated to train *scPrivacy*. It should be noted that in all benchmark scenarios in our study, macro-F1 score was used as the evaluation metric and only query cell types included in multiple institutional datasets were calculated. As shown in Figure 2A and Table S2 in Supporting Information, it can be clearly seen that *scPrivacy* with multiple institutions generally obtained great improvement compared with that of *scPrivacy* with single institution, demonstrating the importance of integrating multiple institutional datasets. To further analyze the results, we compared macro-F1 score on “PBMC-Mereu” and “PBMC-Ding” studies in terms of each cell type, as the two studies shared several common cell types. The results showed that *scPrivacy* achieved a better performance in almost all cell types, further demonstrating the benefits and necessity to integrate multiple institutional datasets (Figure 2B and Table S3 in Supporting Information).

Benchmarking *scPrivacy* with non-privacy-preserving multiple reference integrating methods

Then we benchmarked *scPrivacy* with existing non-privacy-preserving multi-reference based single cell type identification methods, including scmap-cluster (Kiselev et al., 2018), SingleR (Aran et al., 2019), Seurat v3 (Stuart et al., 2019) and mtSC (Duan et al., 2021). In this study, Seurat v3 applied a data-level integration strategy; scmap-cluster and SingleR applied a decision-level integration strategy, and mtSC ap-

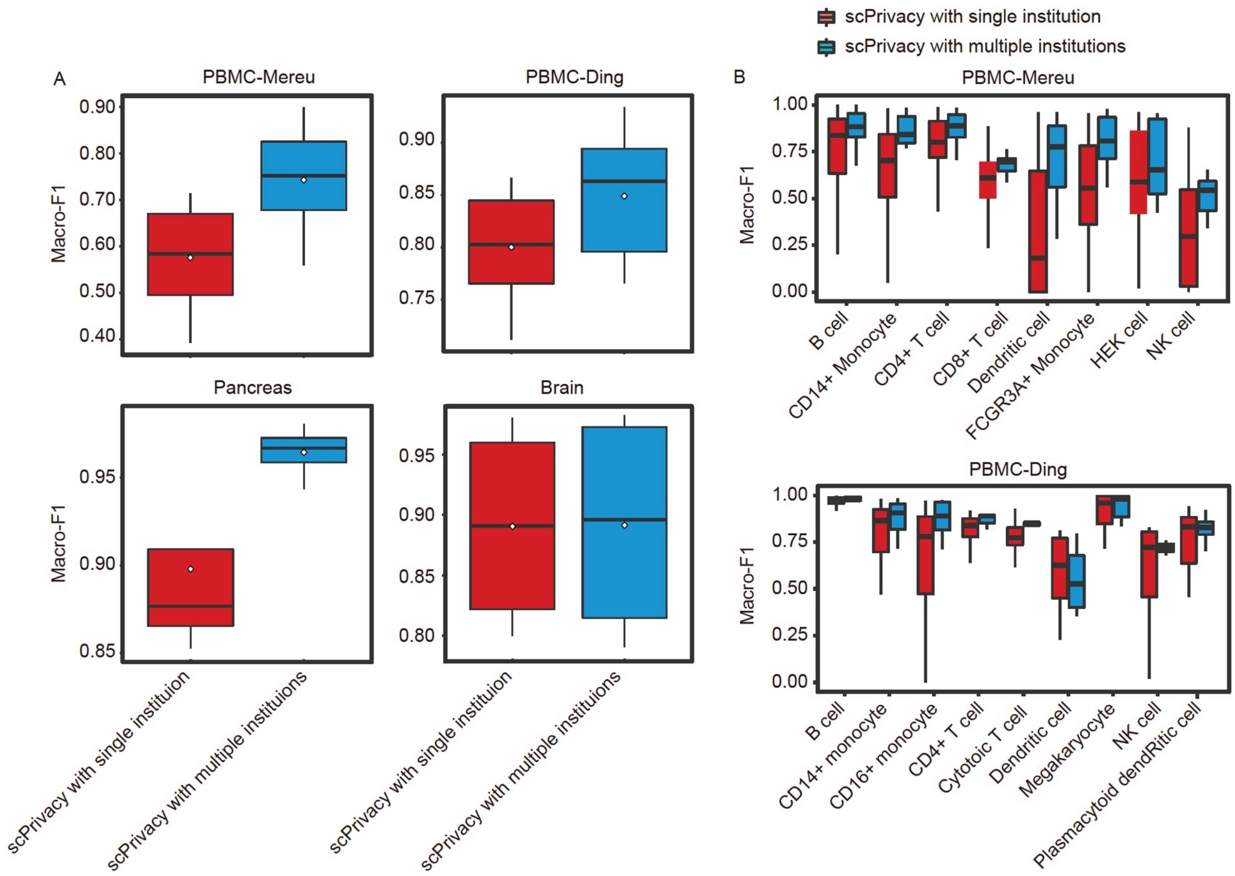


Figure 2 Benchmarking *scPrivacy* with single institution and multiple institutions. A, The macro-F1 scores of *scPrivacy* with single institution and multiple institutions for “PBMC-Mereu”, “PBMC-Ding”, “Brain” and “Pancreas” studies, respectively. The white diamond represents the mean value. B, The macro-F1 of each cell type for *scPrivacy* with single institution and multiple institutions on “PBMC-Ding” and “PBMC-Mereu” studies.

plied an algorithm-level integration strategy (See Methods), which represent the three mainstream integrative analysis strategies for single cell assignment. The above 27 datasets in 4 studies are used as the benchmark datasets here. In this case, each dataset among the multiple datasets was treated as the query, and the others were used to simulate multiple institutional datasets. It should be noted that *scPrivacy* integrated multiple institutional datasets in a data privacy-preserving manner while other multi-reference based methods accessed all datasets to integrate them directly. The results are shown in Figure 3A and Table S4 in Supporting Information. We can see that *scPrivacy* achieved comparable performance to mtSC and performed better than the other methods, while it was trained in a data protection manner. Furthermore, as shown in Figure 3B and Table S5 in Supporting Information, we compared the macro-F1 of each cell type on “PBMC-Ding” and “PBMC-Mereu” studies, and achieved the consistent conclusion that *scPrivacy* achieved the comparable best performance in all four studies as those of mtSC. As a conclusion, *scPrivacy* is able to achieve a comparable best performance to integrate multiple institutional datasets, while keeping the integration in a data privacy-preserving manner.

scPrivacy consumes much less time than most multiple reference integrating methods

As the amount and size of scRNA datasets increasing rapidly, consuming time is an important concern for single cell integration and annotation. Due to the distributed properties of federated learning, the training time of *scPrivacy* only depends on the max training time among the institutions datasets while the training time of other multi-reference based methods are the sum of training time of all institution datasets. Thus, *scPrivacy* obtains a high scalability to deal with the large-scale data integration and model training, which serves a very challenging issue in the building of large-scale atlas level single cell reference data. The following training time and query time comparisons further proved this point (Figure 3C, Tables S6 and S7 in Supporting Information): (1) *scPrivacy* consumes much less time than Seurat v3 and mtSC in the training process. More importantly, the time consumed by *scPrivacy* does not increase exponentially or linearly, since it only depends on the max training time among institutions datasets, indicating its potential ability to handle large-scale datasets. (2) For querying process, *scPrivac*y is also very fast (<1 min for 9,000 query

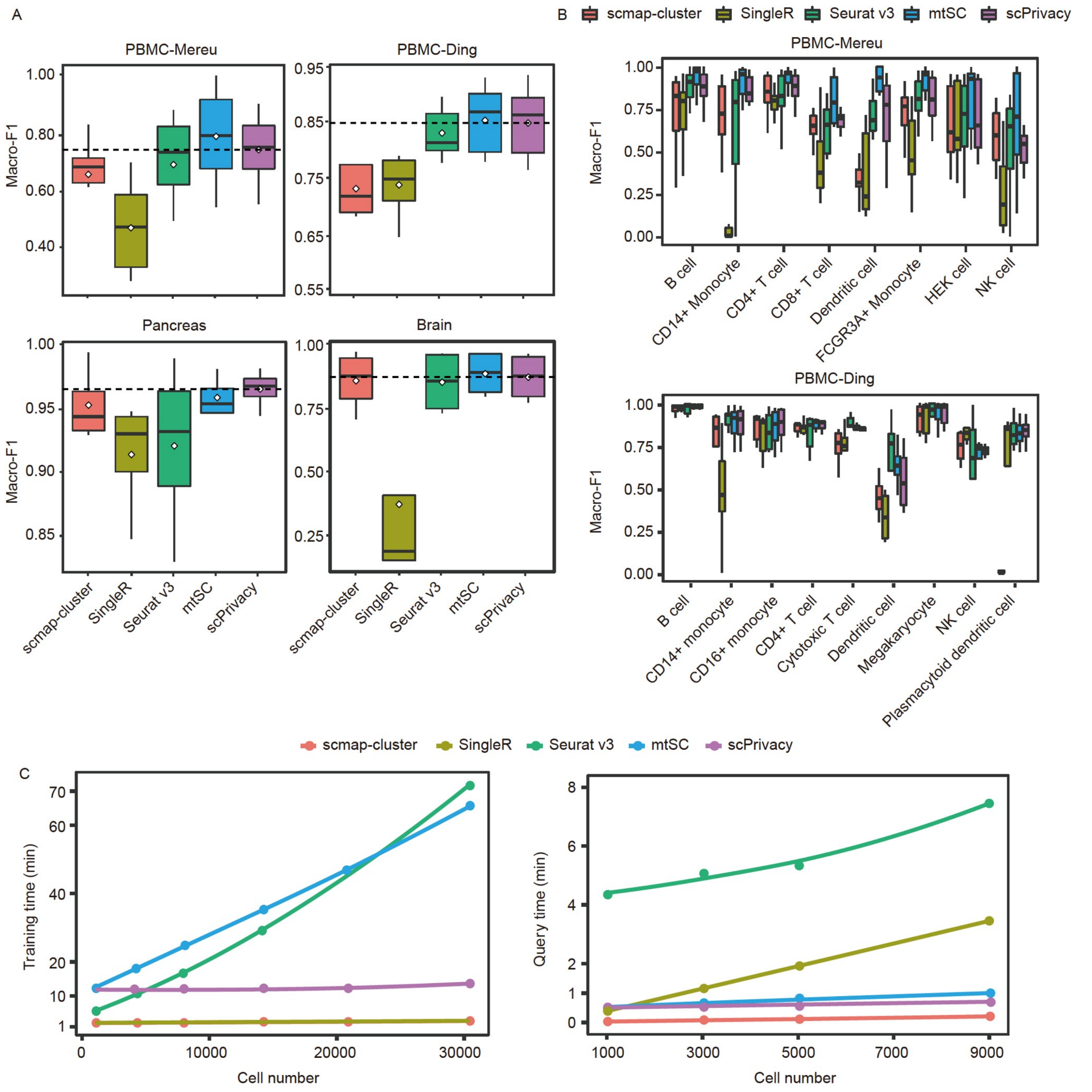


Figure 3 Benchmarking *scPrivacy* with non-privacy-preserving multiple reference integrating methods. A, The macro-F1 scores of *scPrivacy* and other existing non-privacy-preserving multiple reference integrating methods on "PBMC-Mereu", "PBMC-Ding", "Brain" and "Pancreas" studies, respectively. B, The macro-F1 of each cell type for *scPrivacy* and other existing non-privacy-preserving multiple reference integrating methods on "PBMC-Ding" and "PBMC-Mereu" studies. C, Training and query time of *scPrivacy* and other existing non-privacy-preserving multiple reference integrating methods. Solid lines are loess regression fitting (span = 2), implemented with R function geom smooth().

cells) and consumes much less time than all the other methods except for scmap-cluster, since scmap-cluster is the simplest method with substantial expense of accuracy.

Robustness validation of *scPrivacy*

We then validated the robustness of *scPrivacy* using these benchmark data. We firstly investigated the impact of the number of institutions on the performance of *scPrivacy*. We considered each dataset in "PBMC-Ding" as an institution dataset to simulate the scenario. In this test, *scPrivacy* was trained with different numbers of institution datasets, and then the corresponding macro-F1 score was calculated to show the trend of the performance as the number of institution datasets increased. Specifically, each time, we randomly selected one dataset from the 7 datasets as the query dataset and selected 1 to 6 datasets without replacement from the remaining datasets as the institutional reference datasets. This process was repeated 5 times to reduce randomness. As shown in Figure 4A and Table S13 in Supporting Information, we can see that *scPrivacy* generally performs increasingly better as the number of institutions increases. The macro-F1 scores for each dataset as query dataset can be found in Figure S1 in Supporting Information. Specifically, if the performance is relatively lower in the beginning, the improvement becomes more evident as the number of institution datasets increases. Such an improvement obtained by increasing institution datasets is very important for developing an effect and robust cell annotation system in the era of explosive growth of single-cell datasets. Together with the fast training of *scPrivacy*, it is expected to integrate a growing number of institution datasets to obtain an increasingly better cell annotation in a privacy-preserving manner.

Then we explored the influence of similarity metrics for query cell assignment. Another two common similarity metrics i.e., cosine similarity and spearman correlation coefficient were tested here. As shown in Figure 4C, using different similarity metrics has little impact on the results and *scPrivacy* is robust to different similarity metrics. In addition, we explored the influence of the data of an institution completely different from that of other institutions. In our understanding, different tissue datasets are heterogeneous and they are different from each other. To explore the scenario that the data of one institution is completely different from that of other institutions, it is a proper choice to simulate the data of an institution from a different tissue while the others are from the same tissue. To this end, each dataset among PBMC-Ding was treated as the query, and the others and a different tissue dataset (the brain dataset MTG or pancreas dataset Muraro) were used to simulate multiple institutional datasets. As shown in Figure 4B, when the data of one institution is completely different from that of other institutions, the model performance will only decrease slightly.

Finally, we explored the influence of greatly varying data volume across institutions. To simulate the scenario, we selected datasets whose data volumes vary greatly as the institution datasets. In PBMC-Ding datasets, dataset a10Xv2 (9,683 cells) and SM2 (475 cells) were considered as the institutions datasets and the others were treated as the query. In pancreas datasets, dataset Baron (8,562 cells) and Segerstolpe (2,126 cells) were considered as the institutions datasets and the others were treated as the query. As shown in Figure 4D, in general, *scPrivacy* with multiple institutions generally obtained improvement compared with that of *scPrivacy* with single institution, which is consistent with the conclusion in Figure 2. In addition, the results showed that the improvement of the institution with the less data volume is larger than that of the institution with the larger data volume. Taken together, *scPrivacy* is robust to the number of institutions, similarity metrics, data heterogeneity and data volumn, further indicating that it has the ability to handle the complex situation for single cell data integration in real world.

A large-scale simulation of collaborations between multiple hospitals for COVID-19 patient cell annations with *scPrivacy*

In this study, we made a large-scale simulation of a real world scenario that multiple hospitals collaborate together to build an automated cell type annotation system for COVID-19 patients (Figure 5A). It is obvious that the patient data always has privacy issue and hospitals can not share patient data with each other if they are not approved by patients. Zhang et al. (Ren et al., 2021) recently constructed a large-scale PBMC single cell transcriptome atlas which consists of 196 individuals in 5 disease stages from 39 institutes or hospitals. These data are selected to simulate the collaboration between multiple hospoitals for COVID-19 patient cell annotion. In this study, as several integrating methods (such as Seurat v3) can not handle the simulation using all individuals, we randomly selected 15 large-scale individuals datasets (>6,000 cells) satisfying certain criteria (see Methods) (Table S12 in Supporting Information) from different hospitals. The similar benchmark strategies descripted aforementioned were also adopted here. We firstly benchmarked *scPrivacy* with multiple hospitals compared with that of *scPrivacy* with single hospital. As shown in Figure 5B and C and Tables S7, S8 in Supporting Information, *scPrivacy* with multiple hospitals also obtained great improvement in general, especially in terms of cell types which are more difficult to be distinguished compared with that of *scPrivacy* with single hospital, demonstrating the effectiveness of integrating multiple hospital patients datasets. Then we benchmarked *scPrivacy* with existing non-privacy-preserving data integration methods. As shown in Figure 5D, E and Tables S9, S10 in Supporting Information, we can see that *scPrivacy* achieved state-of-the-art performance in all cell types while it was trained in a data protection manner. As a conclusion, this large-scale real world application simiultion also proved that *scPrivacy* is able to handle large-scale data integration issue and achieve state-of-the-art performance to integrate multiple institutional datasets in a data privacy-preserving manner.

DISCUSSION

In this study, we present the first and generalized federated deep metric learning-based single cell type identification

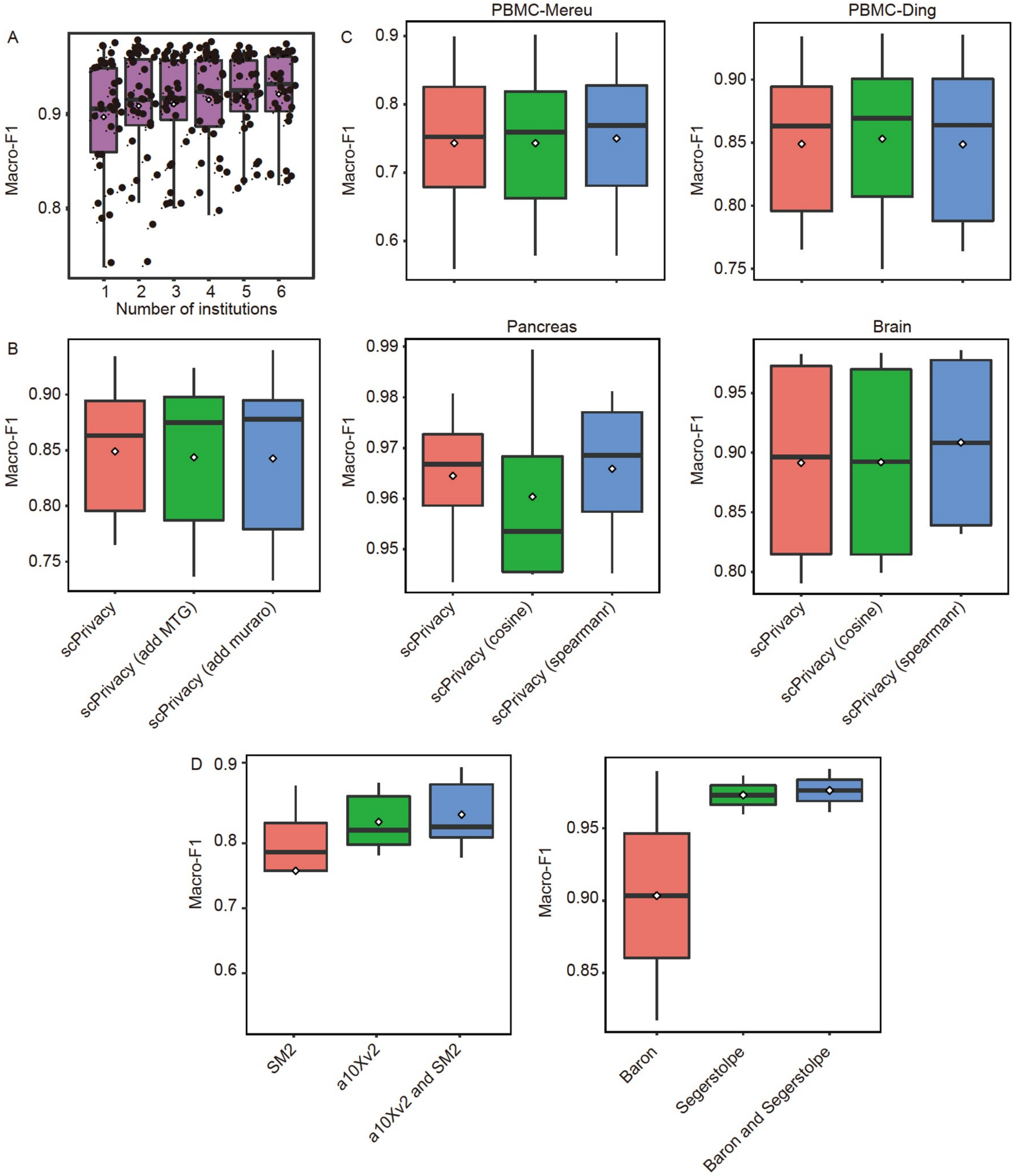


Figure 4 Robustness validation of scPrivacy. A, The performance of *scPrivacy* as the number of institution datasets increases. We considered each dataset in "PBMC-Ding" as an institution dataset to simulate the scenario, and showed the macro-F1 scores for overall datasets. B, The macro-F1 scores of simulation that the data of an institution is from a different tissue while the others are from the same tissue. Each dataset among PBMC-Ding was treated as the query, and the others and a different tissue dataset (the brain dataset MTG or pancreas dataset Muraro) were used to simulate multiple institutional datasets. C, The macro-F1 scores of *scPrivacy* using different similarity metrics including Pearson correlation coefficient, cosine similarity and Spearman correlation coefficient. D, The macro-F1 scores of simulation that the data volumn of institutions vary greatly. In PBMC-Ding datasets, dataset a10Xv2 and SM2 were considered as the institutions datasets and the others were treated as the query. In pancreas datasets, dataset Baron and Segerstolpe were considered as the institutions datasets and the others were treated as the query. The white diamond represents the mean value.

prototype *scPrivacy* to facilitate single cell annotations by integrating multiple institutional datasets in a privacy-preserving manner. As scRNA-seq datasets grow exponentially, multiple institutional datasets can be integrated to build a more comprehensive, effective and robust cell annotation system. Traditional multi-reference based methods faced the problem of data privacy protection. *scPrivacy* solves this issue by federated learning. Specifically, *scPrivacy* trains each institution dataset locally and aggregates encrypted model parameters for all institutions instead of putting raw data of all institutions together to train a model. We evaluated *scPrivacy* on a comprehensive set of publicly available

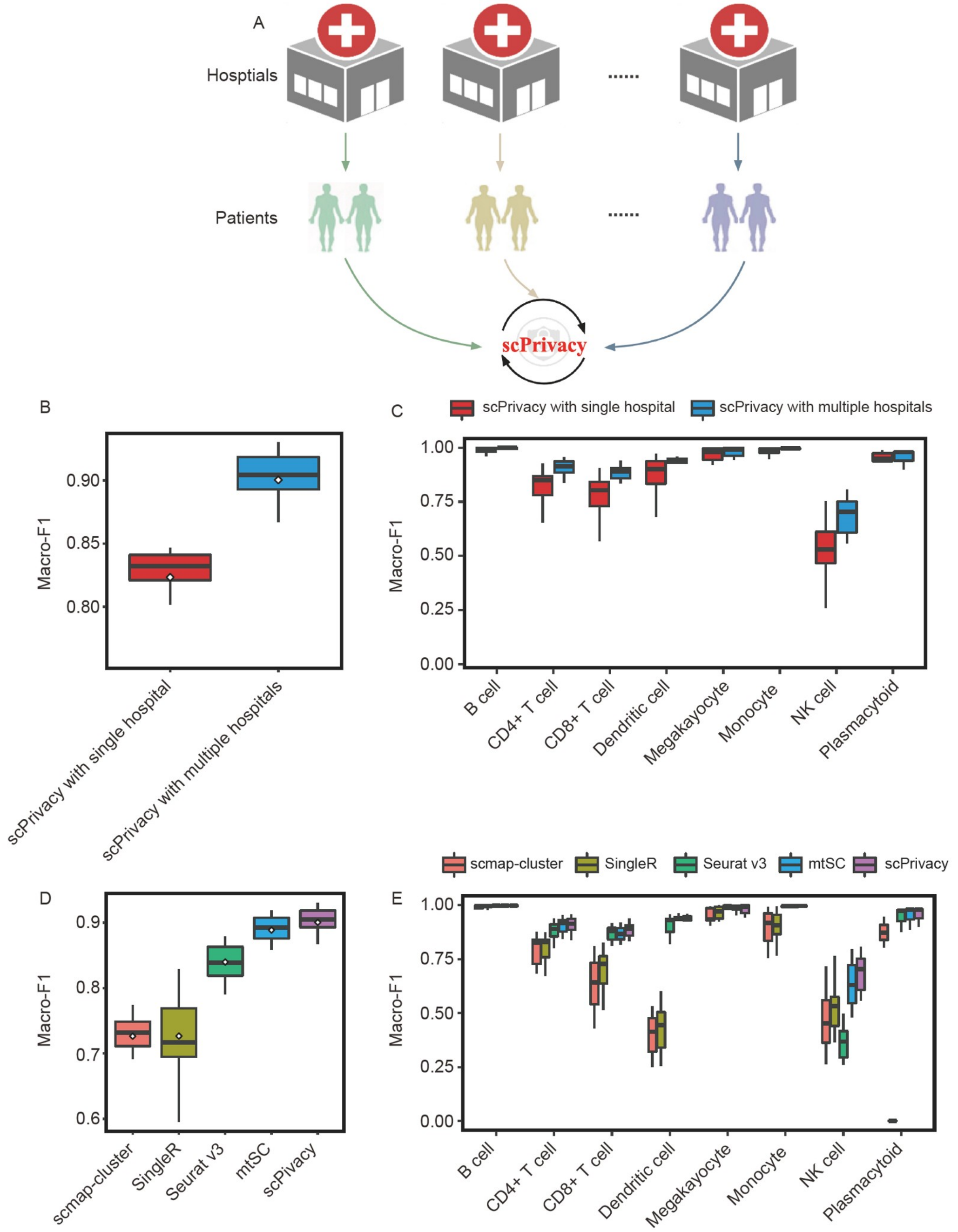


Figure 5 A large-scale real world simulation that multiple hospitals collaborate together to build an automated cell type annotation system for COVID-19 patients with *scPrivacy*. A, The workflow of the simulation. B, The macro-F1 scores of *scPrivacy* with single hospital an d multiple hospitals. C, The macro-F1 of each cell type for *scPrivacy* with single hospital and multiple hospitals. D, The macro-F1 scores of *scPrivacy* and other existing non-privacy-preserving multiple reference integrating methods. E, The macro-F1 of each cell type for *scPrivacy* and other existing non-privacy-preserving multiple reference integrating methods. The white diamond represents the mean value.

benchmark datasets for single cell type identification to stimulate the scenario that the reference datasets are rapidly generated and distributed in multiple institutions, while they are prohibited to be integrated directly or exposed to each other due to the data privacy regulations, and demonstrated its effectiveness for privacy-perserving integration of multiple institutional datasets. In addition, a large-scale real world simulation that multiple hospitals collaborate together to build an automated cell type annotation system for COVID-19 patients with *scPrivacy* was also demonstrated. Collectively, *scPrivacy* is time efficient, performing increasingly better as the number of institution datasets increases, and robust to the number of institutions, similarity metrics, data heterogeneity and data volumn, which is of great potential utility in various real world applications to build single cell atalas in a privacy-preserving way.

In addition to scRNA-seq technology, other single cell omics, such as scATAC-seq and etc., are developing rapidly and vast numbers of datasets will accumulate in the future. Integrating these multi-omics datasets has the potential to provide a more comprehensive picture of basic biological processes. However, privacy issue is still an unavoidable problem for data sharing and integration. *scPrivacy* can be easily extended to other single cell omics for privacy-preserving integration. In addition, the privacy-preserving computing technologies are evolved rapidly, although federated learning is a suitable solution for large-scale privacy-preserving data integration, it still needs a central server to aggregate model parameters from clients. Recently, blockchain-based federated learning, such as swarm learning (Saldanha et al., 2022; Warnat-Herresthal et al., 2021), is developed which does not need the central server, however, its efficiency is waiting to be further evaluated. Nevertheless, a privacy-preserving integration of multiple institutional data globally to build a more comprehensive cell landscape is a challenging issue that must be faced under the global data sharing protection and regulations. We therefore call for attention to such problems and efficient privacy-preserving system for cell annotations are needed to be carefully designed.

METHODS

Benchmark data collection. We evaluated *scPrivacy* on 27 single-cell type identification benchmark datasets and 15 patients datasets which were curated from five studies including three tissues: peripheral blood mononuclear cells (PBMCs) (Ding et al., 2019; Mereu et al., 2020; Ren et al., 2021), the brain (Tasic et al., 2016; Tasic et al., 2018) and the pancreas (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Xin et al., 2016) (Tables S1 and S12 in Supporting Information). For all datasets, cell types with less than 10 cells were removed since they do not contain enough information and are unreliable for subsequent assignment. Cells labeled “alpha.contaminated”, “beta.contaminated”, “gamma.contaminated” and “delta.contaminated” in the dataset generated from Xin et al. (Xin et al., 2016) were removed because they likely corresponded to cells of lower quality. Cells labeled “not applicable” in the dataset generated by Segerstolpe et al. (Segerstolpe et al., 2016) were removed. “L6b”, “Pvalb”, “Sst” and “Vip” cell types in the dataset generated from Tasic et al. (Tasic et al., 2018) were retained to match the names of cell types in the other three brain datasets (Tasic et al., 2016). For “PBMC-Mereu” (Mereu et al., 2020), 12 datasets were used and Smart-seq2-based dataset was excluded, which was too small (Table S1 in Supporting Information).

Data preprocessing. The data preprocessing step of *scPrivacy* consists of three parts: cell quality control, rare cell type filtering and gene expression profile formatting. *scPrivacy* evaluates the cell quality on strict criteria. In particular, the number of genes detected requires >500, the number of unique molecular identifiers induced requires >1,500, and the percentage of mitochondrial genes detected requires <10% among all genes. Only cells satisfying all three criteria are reserved. The quality control of Zhang’s study datasets were processed with the quality control procedures in their paper (Ren et al., 2021). Then, all the datasets were normalized by scaling to 10,000 and then with log (counts+1). Finally, all the datasets are processed into an identical format, i.e., expression profiles with the union of the genes in all institution datasets. The column of the gene in query dataset will be filled with zeros if the gene is not in the gene union of the institution datasets.

The selection of the simulation datasets for collaboration among multiple hospitals in the cell annotations of COVID-19 patient. We filtered the datasets not statisfying following criteria: (1) in particular, the number of cells requires >6,000 to gurantee the large scale of dataset, (2) the number of cell types requires >6 to gurantee the quality of dataset. We randomly selected 15 individuals datasets statisfying criteria in various disease stages from different hospitals and there are 3 individuals datasets avaliable in each disease stage. The details of selected datasets can be found in Table S12 in Supporting Information.

Model learning of *scPrivacy*. In the model learning stage, *scPrivacy* utilizes federated deep metric learning algorithms to train the federated model on multiple institutional datasets in data privacy protection manner. We remove batch effects for datasets by sending the gradients learned from each dataset, then the aggregated model can utilize the information of the same cell type from different datasets so as to uncover the common biological information and remove the batch effect, which is similar to the idea of *mtSC* to remove batch effect (Duan et al., 2021). For a single institution, we use deep metric learning (DML) as the training algorithm and N-pair loss (Sohn, 2016) is used as the loss function. DML is applied to learn an optimal measurement fitting the relationship among cells in the reference dataset. With the measurement learned from DML, cells with the same label become more similar and cells with different labels become more dissimilar. The DML neural network for a single institution contains an input layer, a hidden layer and an output layer. The nodes of input layer equal to the genes of the reference while the nodes of the hidden layer and output layer are 500 and 20, respectively.

The application of the *N*-pair loss consists of two parts: batch construction and calculation. For the batch construction of the *N*-pair loss, $\left\{\left(x_1, x_1^+\right) \cdots , \left(x_N, x_N^+\right)\right\}$ is defined as *N* pairs of cells from *N* different cell types, in which $x_i \neq x_j \ \forall \ i \neq j$. Then, *N* tuples denoted by $\{S_i\}_{i=1}^{N}$ are built from the *N* pairs, where $S_i = \left\{x_i, x_1^+, x_2^+, \dots, x_N^+\right\}$. Here, x_i is the query for S_i, x_i^+ is a positive example, and $x_j^+ (j \neq i)$ are the negative examples. x_i and x_i^+ are two cells of the same cell type, and x_j^+ are the cells with different cell types different from x_i.

The calculation of the N-pair loss can be formulated as follows:

$$L_{N-\text{pair}}\left(\left\{\left(x_i,\ x_i^+\right)\right\}_{i=1}^{N}; f\right) = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+\sum_{j\neq i}\exp\left(f_i^{\mathrm{T}}f_j^+ - f_i^{\mathrm{T}}f_i^+\right)\right) \quad (1)$$

in which $f(\cdot;\ \theta)$ is an embedding kernel defined by a deep neural network, f_i and f_i^+ are embedding vectors of two cells of the same cell type and f_j^+ are embedding vectors of cells whose cell types are different from x_i.

scPrivacy extends DML to a federated learning framework. Define N institutions $\{F_1,...,F_N\}$ and their respective data $\{D_1,...,D_N\}$. A federated learning framework allows institutions to collaboratively train a model and no institution F_i expose its data D_i to others.

As shown in Figure 1, in *scPrivacy*, institutions datasets learn a federated model collaboratively with the help of a server. To train a federated model, our training process can be divided into the following four steps (Yang et al., 2019):

• Step 1: Each institution trains its model on its own dataset with DML;

• Step 2: Institutions encrypt their own model parameters and send encrypted parameters to server;

• Step 3: Server performs secure aggregation for encrypted parameters;

• Step 4: Server sends back the aggregated model parameters to institutions and institutions update their models with the decrypted aggregated model parameters.

In our implementation of *scPrivacy*, we used the Crypten, a ML framework built on PyTorch (Paszke et al., 2019) to implement encryption of model parameters and their operations in secure aggregation with secure multi-party computation (Yao, 1982) which is adopted as an encryption technique here. At the central server, we adapted widely-used FedAvg (McMahan et al., 2016) algorithm to aggregate model parameters. The basic idea of the algorithm to aggregate model parameters is to average model parameters (weight parameters w and bias parameters b) of the neural network models for institutions. Institutions update their models by replacing model parameters with the decrypted aggregated model parameters.

Model parameters of *scPrivacy*. The neural network model was implemented with PyTorch. We use Adam optimizer to update parameters of the network via backpropagation. The learning rate was set to 0.0005. The number of training epochs was set to 300. The L2 regularization rate was set to 0.05.

Query cells assignment. First, the query dataset was scaled to 10,000 and normalized with log(counts+1). The column of the gene will be filled with zeros if the gene is not in the gene union of the institution datasets. Next, the query cells were transformed by the federated model to the same embedding space as that of the transformed institution datasets. Then, the transformed query dataset was assigned to proper cell types by comparing with cell type landmarks of transformed institution datasets. Specifically, for each transformed institution dataset, *scPrivacy* carried out a cell search by measuring similarity between transformed query cells and cell type landmarks of the transformed institution datasets. Pearson correlation coefficient was adopted to calculate similarity as in our previous study (Duan et al., 2020). Finally, the query cells obtained the predicted cell types with the highest similarity among all landmarks of the transformed institution datasets.

Benchmarking existing tools. To evaluate the performance of *scPrivacy*, four classical non-privacy-preserving data-integration methods were compared: Seurat v3 (Stuart et al., 2019), scmap-cluster (Kiselev et al., 2018), SingleR (Aran et al., 2019) and mtSC (Duan et al., 2021). Seurat v3 applied a data-level integration strategy, where multiple datasets are integrated into one dataset directly; scmap-cluster and SingleR applied a decision-level integration method, where the final assignment results are ensembled by individual assignment results; and mtSC applied an algorithm-level integration strategy, where efficient algorithms are designed for model integration, while keeping datasets separately. However, all these methods are needed to access the individual institutional data directly. For a fair comparsion purpose, all these methods were allowed to access all institutional datasets directly while *scPrivacy* integrated multiple institutional datasets in a data privacy-preserving manner. Each dataset among the multiple datasets in a study was treated as the query, and the others were used to simulate multiple institutional datasets. For scmap-cluster, "threshold = 0" was set to not assign query cells to "unassigned". For SingleR, as the fine-tuning process of SingleR is extremely time consuming, "fine.tune = FALSE" was set. For Seurat v3, all parameters were the defaults. For mtSC, all parameters were the defaults. In all tests, Seurat v3, scmap-cluster and SingleR and were trained and tested with CPU Intel Xeon Platinum 8165 2.3–3.7 GHz. The deep learning based methods mtSC and *scPrivacy* were trained with GPU 1080Ti and tested with the same CPU as the other methods. To further prove the effectiveness of *scPrivacy*, we also benchmarked *scPrivacy* with the decision-level strategy (Figure S2 in Supporting Information).

Evaluation criteria. The macro-F1 score was used as evaluation criteria. Despite integration tasks, we calculated the macro-F1 score for each query dataset just like the calculation for no-integration tasks. First, the precision and recall of each cell type were calculated. Then, the macro-F1 score was calculated as listed below:

$$\text{macro}-\text{F1}=\frac{1}{N}\sum_{i=1}^{N}\frac{2*\text{Precision}_i*\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i}$$

in which N denotes the number of cell types in a dataset and Precision_i and Recall_i are the precision and recall of the i-th cell type in the dataset.

Data availability

The 27 single-cell type identification benchmark datasets and 15 patients datasets were curated from five studies including three tissues: peripheral blood mononuclear cells (PBMCs) (Ding et al., 2019; Mereu et al., 2020; Ren et al., 2021), the brain (Tasic et al., 2016; Tasic et al., 2018) and the pancreas (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Xin et al., 2016) (Tables S1 and S12 in Supporting Information). The four pancreas datasets (Baron et al., 2016; Muraro et al., 2016; Segerstolpe et al., 2016; Xin et al., 2016) and one of the brain datasets (Tasic et al., 2018) were collected in previous work of scmap (Kiselev et al., 2018) (https://hemberg-lab.github.io/scRNA.seq.datasets), and the other three brain datasets and seven datasets in "PBMC-Ding" (Ding et al., 2019) were curated from the following benchmark study (Abdelaal et al., 2019) (https://doi.org/10.5281/zenodo.3357167). The 12 datasets in "PBMC-Mereu" (Mereu et al., 2020) were collected from GSE133549, and the corresponding RData file can be downloaded in https://www.dropbox.com/s/i8mwmyymchx8mn8/sce.all_classified.technologies.RData?dl=0. All these datasets were converted into Bioconductor SingleCellExperiment (http://bioconductor.org/packages/SingleCellExperiment) class objects. The 15 datasets of COVID-19 patients were collected from GSE158055 (Table S12 in Supporting Information).

Code availability

scPrivacy is developed as a python package for simulations, which is available at https://github.com/bm2-lab/scPrivacy.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *This work was supported by the National Key Research and Development Program of China (2021YFF1200900, 2021YFF1201200), the National Natural Science Foundation of China (31970638, 61572361), the Shanghai Artificial Intelligence Technology Standard Project (19DZ2200900), the Shanghai Shuguang Scholars Project, WeBank Scholars Project and the Fundamental Research Funds for the Central Universities.*

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol 20, 194.

Acar, A., Aksu, H., Uluagac, A.S., and Conti, M. (2019). A survey on homomorphic encryption schemes. ACM Comput Surv 51, 1–35.

Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol 20, 163–172.

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 Syst 3, 346–360.e4.

Benefield, H., Ashkanazi, G., and Rozensky, R.H. (2006). Communication and records: hippa issues when working in health care settings. Prof Psychol-Res Pract 37, 273–277.

Byrd, J.B., Greene, A.C., Prasad, D.V., Jiang, X., and Greene, C.S. (2020). Responsible, practical genomic data sharing that accelerates research. Nat Rev Genet 21, 615–629.

Chen, S., Luo, Y., Gao, H., Li, F., Chen, Y., Li, J., You, R., Hao, M., Bian, H., Xi, X., et al. (2022a). hECA: the cell-centric assembly of a cell atlas. iScience 25, 104318.

Chen, S., Luo, Y., Gao, H., Li, F., Li, J., Chen, Y., You, R., Lv, H., Hua, K., Jiang, R., et al. (2022b). Toward a unified information framework for cell atlas assembly. Natl Sci Rev 9, nwab179.

Chen, S., Xue, D., Chuai, G., Yang, Q., and Liu, Q. (2021). FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. Bioinformatics 36, 5492–5498.

Ding, J., Adiconis, X., Simmons, S.K., Kowalczyk, M.S., Hession, C.C., Marjanovic, N.D., Hughes, T.K., Wadsworth, M.H., Burks, T., Nguyen, L.T., Kwon, J.Y.H., Barak, B., Ge, W., Kedaigle, A.J., Carroll, S., Li, S., Hacohen, N., Rozenblatt-Rosen, O., Shalek, A.K., Villani, A.-C., Regev, A., and Levin, J.Z. (2019). Systematic comparative analysis of single cell RNA-sequencing methods. bioRxiv, 632216.

Domínguez Conde, C., Xu, C., Jarvis, L.B., Rainbow, D.B., Wells, S.B., Gomes, T., Howlett, S.K., Suchanek, O., Polanski, K., King, H.W., et al. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science 376, eabl5197.

Duan, B., Chen, S., Chen, X., Zhu, C., Tang, C., Wang, S., Gao, Y., Fu, S., and Liu, Q. (2021). Integrating multiple references for single-cell assignment. Nucl Acids Res 49, e80.

Duan, B., Zhu, C., Chuai, G., Tang, C., Chen, X., Chen, S., Fu, S., Li, G., and Liu, Q. (2020). Learning for single-cell assignment. Sci Adv 6, eabd0855.

Elmentaite, R., Ross, A.D.B., Roberts, K., James, K.R., Ortmann, D., Gomes, T., Nayak, K., Tuck, L., Pritchard, S., Bayraktar, O.A., et al. (2020). Single-cell sequencing of developing human gut reveals transcriptional links to childhood crohn's disease. Dev Cell 55, 771–783.e5.

Eraslan, G., Drokhlyansky, E., Anand, S., Fiskin, E., Subramanian, A., Slyper, M., Wang, J., Van Wittenberghe, N., Rouhana, J.M., Waldman, J., et al. (2022). Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. Science 376, eabl4290.

Guan, Y.N., Li, Y., Roosan, M., and Jing, Q. (2021). Single-cell transcriptomics of murine mural cells reveals cellular heterogeneity. Sci China Life Sci 64, 1077–1086.

Halamka, J.D., and Tripathi, M. (2017). The HITECH era in retrospect. N Engl J Med 377, 907–909.

Jiang, H., Zhang, H., and Zhang, X. (2021). Single-cell genomic profile-based analysis of tissue differentiation in colorectal cancer. Sci China Life Sci 64, 1311–1325.

Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat Methods 15, 359–362.

Li, C., Liu, B., Kang, B., Liu, Z., Liu, Y., Chen, C., Ren, X., and Zhang, Z.

(2020). SciBet as a portable and fast single cell type identifier. Nat Commun 11, 1818.

Liu, J., Li, J., Wang, H., and Yan, J. (2020). Application of deep learning in genomics. Sci China Life Sci 63, 1860–1878.

Liu, Z., and Zhang, Z. (2022). Mapping cell types across human tissues. Science 376, 695–696.

Lotfollahi, M., Naghipourfar, M., Luecken, M.D., Khajavi, M., Büttner, M., Wagenstetter, M., Avsec, Ž., Gayoso, A., Yosef, N., Interlandi, M., et al. (2022). Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol 40, 121–130.

Ma, F., and Pellegrini, M. (2020). ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics 36, 533–538.

McKeen, F., Alexandrovich, I., Anati, I., Caspi, D., Johnson, S., Leslie-Hurd, R., and Rozas, C. (2016). Intel® Software Guard Extensions (Intel® SGX) Support for Dynamic Memory Management Inside an Enclave. In Proceedings of the Hardware and Architectural Support for Security and Privacy 2016 on - HASP 2016, pp. 1–9.

McMahan, H.B., Moore, E., Ramage, D., and Hampson, S. (2016). Communication-efficient learning of deep networks from decentralized data. arXiv preprint, .

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, E., Grün, D., Lau, J.K., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol 38, 747–755.

Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., et al. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Syst 3, 385–394.e3.

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A.M.P., George, N., Fexova, S., Fonseca, N.A., Füllgrabe, A., Green, M., Huang, N., et al. (2019). Expression Atlas update: from tissues to single cells. Nucl Acids Res 48, D77–D83.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. (2019). PyTorch: an imperative style, high-performance deep learning library. Paper presented at: Advances in Neural Information Processing Systems. (New York: ACM), pp. 8026–8037.

Plass, M., Solana, J., Wolf, F.A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F.J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science 360.

Politou, E., Alepis, E., and Patsakis, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. J Cybersecur 4.

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.

Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. Cell 184, 1895–1913.e19.

Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., et al. (2020). The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. Cell 181, 236–249.

Saldanha, O.L., Quirke, P., West, N.P., James, J.A., Loughrey, M.B., Grabsch, H.I., Salto-Tellez, M., Alwers, E., Cifci, D., Ghaffari Laleh, N., et al. (2022). Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat Med 28, 1232–1239.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.M., Andréasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. Cell Metab 24, 593–607.

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.

Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. Adv Neur In 29.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck Iii, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell 177, 1888–1902. e21.

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.E., Botting, R.A., Stephenson, E., Engelbert, J., Tuong, Z.K., et al. (2022). Mapping the developing human immune system across organs. Science 376.

Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., et al. (2022). The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. Science 376, eabl4896.

Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci 19, 335–346.

Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al. (2018). Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72–78.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature 587, 619–625.

Warnat-Herresthal, S., Schultze, H., Shastry, K.L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., Händler, K., Pickkers, P., Aziz, N.A., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. Nature 594, 265–270.

Winnubst, J., and Arber, S. (2021). A census of cell types in the brain's motor cortex. Nature 598, 33–34.

Xie, X., Cheng, X., Wang, G., Zhang, B., Liu, M., Chen, L., Cheng, H., Hao, S., Zhou, J., Zhu, P., et al. (2021). Single-cell transcriptomes of peripheral blood cells indicate and elucidate severity of COVID-19. Sci China Life Sci 64, 1634–1644.

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A.J., Yancopoulos, G.D., Lin, C., and Gromada, J. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab 24, 608–615.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) 10, 1–19.

Yao, A.C. (1982). Protocols for secure computations. In: Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science.

Zhang, Y., and Yang, Q. (2018). An overview of multi-task learning. Natl Sci Rev 5, 30–43.

Zhao, Y., Wang, T., Liu, Z., Ke, Y., Li, R., Chen, H., You, Y., Wu, G., Cao, S., Du, Z., et al. (2022). Single-cell transcriptomics of immune cells in lymph nodes reveals their composition and alterations in functional dynamics during the early stages of bubonic plague. Sci China Life Sci, doi: 10.1007/s11427-021-2119-5.

SUPPORTING INFORMATION

The supporting information is available online at https://doi.org/10.1007/s11427-022-2224-4. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.