PPPCT- Privacy-Preserving Framework for Parallel Clustering Transcriptomics Data: Supplementary document

1. RELATED WORK

We investigate the related work from two separate perspectives: 1) approaches for Privacy-preserving of scRNA-seq clustering, 2) approaches for scRNA-seq clustering.

A. Privacy-preserving scRNA-seq clustering

Some recent works ([1–4]) proposed a privacy-preserving semi-supervised mechanism for scRNA-seq clustering. Using semi-supervised clustering techniques in this domain has some drawbacks: I) data labels are limited, II) potential bias in sample labeling, III) high computation time to train the semi-supervised model, and IV) less scalability for managing various numbers of genes and cells. Bozdemir *et al.* ([5]) proposed a privacy-preserving mechanism based on multi-party computation. Multi-party computation provides communication overhead. Thus, an increase in convergence time. Moreover, their approach cannot cluster hundreds of thousands of cells. Mohassel *et al.* ([6]) proposed a private k-means algorithm based on two-party computation and oblivious transfer. Their approach requires a high computation power, as they need to change the operations to boolean gates, which requires a lot of communication between collaborating parties. This results in late convergence in k-means and, accordingly, high computation time. Gheid *et al.* ([7]) and Jaschke *et al.* ([8]) tried to resolve the privacy issue of k-means clustering targeting non-genomic data. Cao *et al.* ([9]) proposed a private framework based on Intel Software Guard eXtensions (SGX) for the scRNA-seq dataset. They combined autoencoders and k-means for dimensionality reduction and clustering, respectively. Although their work is not computationally expensive, they are not employing parallel computing, thus increasing computation time and losing scalability. Moreover, they did not run their experiments on real-world human datasets to see how efficient their work is in terms of ARI, RI, HI, and computation time.

B. scRNA-seq clustering

Cell types are largely unknown in most scRNA-seq studies ([10]). As such, researchers generally employ unsupervised clustering methods to group cells into subsets. Based on the clustering results, they can characterize and determine cell types ([11]). Given the high dimensionality and complex relationships in scRNA-seq data, a widely adopted approach involves combining dimensionality reduction techniques with classic clustering algorithms. Common combinations include manifold learning approaches such as t-SNE plus k-means ([12]), and Principal Component Analysis (PCA) plus hierarchical clustering ([13]). Dirichlet Mixture Model (DMM) ([14]) can work fine for scRNA-seq clustering as the discrete counting information in the Unique Molecular Identifiers (UMI) matrix can be directly modeled through multinomial distribution and conjugate prior likelihood pairs, resulting in efficient inference. Parallel clustering on the Dirichlet Process Mixture Model (ParaDPMM) ([15]) is an example of using the Dirichlet process for clustering single-cell data. SC3 ([16]) introduced a semi-supervised consensus clustering, where subsets of the dataset are selected, then using a consensus matrix, a support vector machine is trained. However, the convergence time is high for high-dimensional and big datasets. Seurat ([17]), first constructs a k nearest-neighbor graph using the Euclidean distance in PCA space, then reassigns the edge weights between any two cells based on the shared overlap in their local neighborhoods. Finally, it cuts the graph into clusters using the Louvain algorithm, which optimizes the modularity between subgroups. Seurat is less computationally intensive than SC3. However, it does not provide high-quality clustering. SIMLR ([18]) and CIDR ([19]) combine modified PCA with hierarchical clustering and a modified t-SNE with k-means, respectively. Wei *et al.* ([20]) introduced an accurate and fast clustering method to find the optimal number of clusters and used a combination of the Uniform Manifold Approximation and Projection (UMAP) technique with spectral clustering to locate the optimal labels of data points. Even though they found a new

way to find the optimal number of clusters in scRNA-seq datasets, the quality of their clustering in real-world datasets is similar to k-means. Recent applications of DMM to single-cell analysis have achieved good results ([21, 22]) in comparison to other approaches. However, there are still some challenges in this field: (1) clustering quality needs to be improved, (2) computational overhead should be decreased, and (3) privacy of patient data should be preserved.

2. INTEL SOFTWARE GUARD EXTENSION (SGX)

SGX are security instructions built into some Intel Central Processing Units (CPUs). With SGX, developers can divide a computer's memory into enclaves, which are private, predefined areas in memory that can better protect sensitive information ([23]). These areas are encrypted by CPU native instructions. Enclaves are essentially stateless ([24]): they are destroyed when the system goes to sleep, when the application exits, and when the application explicitly destroys them. When an enclave is destroyed, all of its contents are lost. To preserve the information stored in an enclave, it must be encrypted and sent outside the enclave to untrusted memory. The sealing process encrypts data in the enclave using an encryption key derived from the CPU ([25]). This encrypted data block can only be decrypted on the same system where it was encrypted unless attestation is used. As part of the attestation, the remote enclave proves the following to the local enclave: its identity, that it has not been tampered with, and that it is running on a genuine platform with Intel SGX enabled. The local enclave can safely send secrets to the remote enclave at that point. Attestation is helpful in preserving the privacy and confidentiality of the data in a cloud environment. The successful result of attestation will offer a protected channel between two local/remote enclaves with a guarantee of confidentiality, integrity, and replay protection.

3. DATA TUNING

This is a very important part of changing the data into a digestible format for PPPCT. Once we are done with dimensionality reduction with NMF, we do a logarithm transform of matrix W based on algorithm 3 in the main paper. Figure S1 shows the output of this transformation gradually after 5 epochs.

REFERENCES

1. T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.
2. M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, *et al.*, "Mapping single-cell data to reference atlases by transfer learning," *Nature biotechnology*, vol. 40, no. 1, pp. 121–130, 2022.
3. J. B. Byrd, A. C. Greene, D. V. Prasad, X. Jiang, and C. S. Greene, "Responsible, practical genomic data sharing that accelerates research," *Nature Reviews Genetics*, vol. 21, no. 10, pp. 615–629, 2020.
4. S. Chen, B. Duan, C. Zhu, C. Tang, S. Wang, Y. Gao, S. Fu, L. Fan, Q. Yang, and Q. Liu, "Privacy-preserving integration of multiple institutional data for single-cell type identification with scprivacy," *Science China Life Sciences*, pp. 1–13, 2022.
5. B. Bozdemir, S. Canard, O. Ermis, H. Möllering, M. Önen, and T. Schneider, "Privacy-preserving density-based clustering," in *Proceedings of the ACM Asia Conference on Computer and Communications Security*, pp. 658–671, 2021.
6. P. Mohassel, M. Rosulek, and N. Trieu, "Practical privacy-preserving k-means clustering," *Cryptology ePrint Archive*, 2019.
7. Z. Gheid and Y. Challal, "Efficient and privacy-preserving k-means clustering for big data mining," in *2016 IEEE Trustcom/BigDataSE/ISPA*, pp. 791–798, IEEE, 2016.
8. A. Jäschke and F. Armknecht, "Unsupervised machine learning on encrypted data," in *Selected Areas in Cryptography–SAC 2018: 25th International Conference, Calgary, AB, Canada, August 15–17, 2018, Revised Selected Papers*, pp. 453–478, Springer, 2019.
9. Q. Cao, "Privacy-preserving clustering of single-cell rna sequencing data on intel sgx," 2021.
10. X. Lin, H. Liu, Z. Wei, S. B. Roy, and N. Gao, "An active learning approach for clustering single-cell rna-seq data," *Laboratory Investigation*, vol. 102, no. 3, pp. 227–235, 2022.
11. V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in unsupervised clustering of single-cell rna-seq data," *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282, 2019.
12. D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and

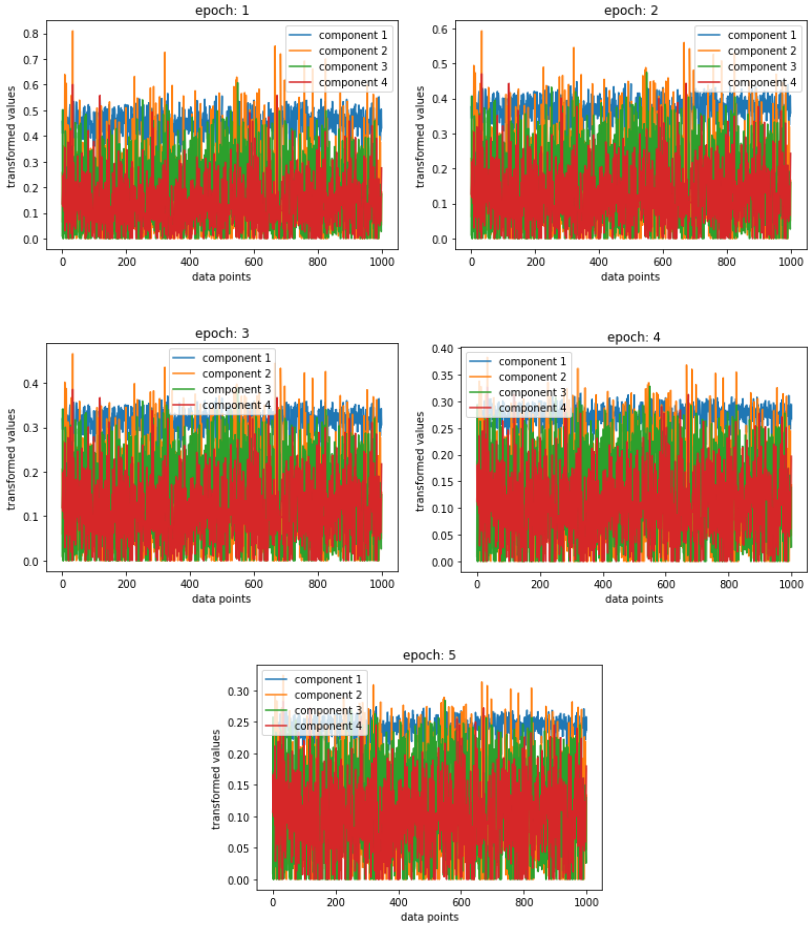

Fig. S1. Results of running data tuning on the first 1000 cells from S-Set after NMF transformation.

A. Van Oudenaarden, "Single-cell messenger rna sequencing reveals rare intestinal cell types," *Nature*, vol. 525, no. 7568, pp. 251–255, 2015.

13. C. Yau *et al.*, "pcareduce: hierarchical clustering of single cell transcriptional profiles," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–11, 2016.
14. Y. Li, E. Schofield, and M. Gönen, "A tutorial on dirichlet process mixture modeling," *Journal of mathematical psychology*, vol. 91, pp. 128–144, 2019.
15. T. Duan, J. P. Pinto, and X. Xie, "Parallel clustering of single cell transcriptomic data with split-merge sampling on dirichlet process mixtures," *Bioinformatics*, vol. 35, no. 6, pp. 953–961, 2019.
16. V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, *et al.*, "Sc3: consensus clustering of single-cell rna-seq data," *Nature methods*, vol. 14, no. 5, pp. 483–486, 2017.
17. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nature biotechnology*, vol. 33, no. 5, pp. 495–502, 2015.
18. B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, and S. Batzoglou, "Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning," *Nature methods*, vol. 14, no. 4, pp. 414–416, 2017.
19. P. Lin, M. Troup, and J. W. Ho, "Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data," *Genome biology*, vol. 18, no. 1, pp. 1–11, 2017.
20. N. Wei, Y. Nie, L. Liu, X. Zheng, and H.-J. Wu, "Secuer: Ultrafast, scalable and accurate clustering of single-cell rna-seq data," *PLOS Computational Biology*, vol. 18, no. 12, p. e1010753, 2022.

21. D. A. duVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda, "Celltree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data," *BMC bioinformatics*, vol. 17, no. 1, pp. 1–17, 2016.
22. Z. Sun, T. Wang, K. Deng, X.-F. Wang, R. Lafyatis, Y. Ding, M. Hu, and W. Chen, "Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data," *Bioinformatics*, vol. 34, no. 1, pp. 139–146, 2018.
23. F. McKeen, I. Alexandrovich, I. Anati, D. Caspi, S. Johnson, R. Leslie-Hurd, and C. Rozas, "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*, pp. 1–9, 2016.
24. V. Costan and S. Devadas, "Intel sgx explained," *Cryptology ePrint Archive*, 2016.
25. K. W. Ahmed, M. M. Al Aziz, M. N. Sadat, D. Alhadidi, and N. Mohammed, "Nearest neighbour search over encrypted data using intel sgx," *Journal of Information Security and Applications*, vol. 54, p. 102579, 2020.