HUMAN GENETICS

PPML-Omics: A privacy-preserving federated machine learning method protects patients' privacy in omic data

Juexiao Zhou[1,2]†, Siyuan Chen[1,2]†, Yulian Wu[1,2]†, Haoyang Li[1,2], Bin Zhang[1,2], Longxi Zhou[1,2], Yan Hu[1], Zihang Xiang[1], Zhongxiao Li[1,2], Ningning Chen[1,2], Wenkai Han[1,2], Chencheng Xu[1,2], Di Wang[1,2]*, Xin Gao[1,2]*

Modern machine learning models toward various tasks with omic data analysis give rise to threats of privacy leakage of patients involved in those datasets. Here, we proposed a secure and privacy-preserving machine learning method (PPML-Omics) by designing a decentralized differential private federated learning algorithm. We applied PPML-Omics to analyze data from three sequencing technologies and addressed the privacy concern in three major tasks of omic data under three representative deep learning models. We examined privacy breaches in depth through privacy attack experiments and demonstrated that PPML-Omics could protect patients' privacy. In each of these applications, PPML-Omics was able to outperform methods of comparison under the same level of privacy guarantee, demonstrating the versatility of the method in simultaneously balancing the privacy-preserving capability and utility in omic data analysis. Furthermore, we gave the theoretical proof of the privacy-preserving capability of PPML-Omics, suggesting the first mathematically guaranteed method with robust and generalizable empirical performance in protecting patients' privacy in omic data.

INTRODUCTION

Individual privacy in biology and biomedicine is emerging as a big concern (*1*) with the development of biomedical data science in recent years. A deluge of genetic data from millions of individuals has been generated from massive research projects in the past few decades, such as The Cancer Genome Atlas (TCGA) (*2*), the 100,000 Genome Project (*3*), and the Earth BioGenome Project (EBP) (*4*) from high-throughput sequencing platforms (*5*). Those datasets may lead to potential leakage of genetic information and privacy concerns on ethical problems like genetic discrimination (*6*). Consequently, a large amount of potentially private genetic information from modern multi-modal sequencing platforms, including bulk RNA sequencing (RNA-seq) (*7*), single-cell RNA-seq (scRNA-seq) (*8*), and spatial transcriptomics (*9*) might also be exposed as more and more data are being published (section S1.1).

In addition to the data itself, another risk factor to data privacy is the wide scope of applications of machine learning (ML), especially deep learning, which evolves rapidly by taking advantage of large datasets. A massive number of models and applications, which require training on a large and diverse dataset (either public data or in-house data), are being created, shared, and applied to various areas, such as genomics (*10*), medical imaging (*11*), and health care (*12*). Nevertheless, individual privacy is being exposed to high risk, leading to the previously unidentified concern of privacy in modern artificial intelligence (AI) (*13*). As shown in Fig. 1A, typically, sensitive data may exist in a distributed manner, where the data owners do not want to share the raw data for privacy reasons, while the aggregators want to access enough data to improve model utility. To balance the needs of both, we need an ML framework for distributed data that can balance utility and privacy-preserving capabilities: a secure and privacy-preserving ML (PPML) method.

To alleviate the leakage of privacy, the most commonly used strategy is the anonymization or pseudonymization of sensitive data before transmitting it to the data-sharing center (*14*). Unfortunately, recent studies showed that anonymization was insufficient for reidentification attacks (*15*) and linking attacks (*16*). To overcome the shortness of centralized data sharing and model training, federated learning (FL) was proposed in 2017 as a data-private collaborative learning method (*17*). The collaborating institutions train an ML model with their own data in parallel and send the model updates to the central server, which can aggregate all model updates into a consensus model without accessing the raw data. Nevertheless, the distributed nature of FL gives rise to previously unidentified threats of privacy leakage caused by potentially malicious participants (*18*–*21*) such as data poisoning attacks (*22*), membership inference attacks (*23*, *24*), source inference attacks (SIA) (*25*), and data reconstruction attack (*26*). Hence, exposing the trained model to a non-trusted user may also cause privacy leakage (*27*).

To further strengthen FL's privacy guarantee to preserve privacy, additional privacy-enhancing modules are required. Within the most extensively studied field, multi-party computation (MPC) or multi-party homomorphic encryption (MHE) frameworks use cryptographic techniques to protect the data while enabling the training of ML models with perfect accuracy. These techniques have been used to secure FL training (*28*–*30*) and achieve stronger privacy protection at the expense of computational efficiency, which might be difficult to satisfy for some clients in practice. MPC incurs a high network-communication overhead and is difficult to scale to a large number of clients, while MHE introduces high storage, computational overheads, and a single point of failure in the standard centralized setup, where one server receives all encrypted datasets to secure federated computation (*31*). Blockchain is also used to secure FL training such as swarm learning

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. [2]Computational Bioscience Research Center, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia.
*Corresponding author. Email: di.wang@kaust.edu.sa (D.W.); xin.gao@kaust.edu.sa (X.G.)
†These authors contributed equally to this work.

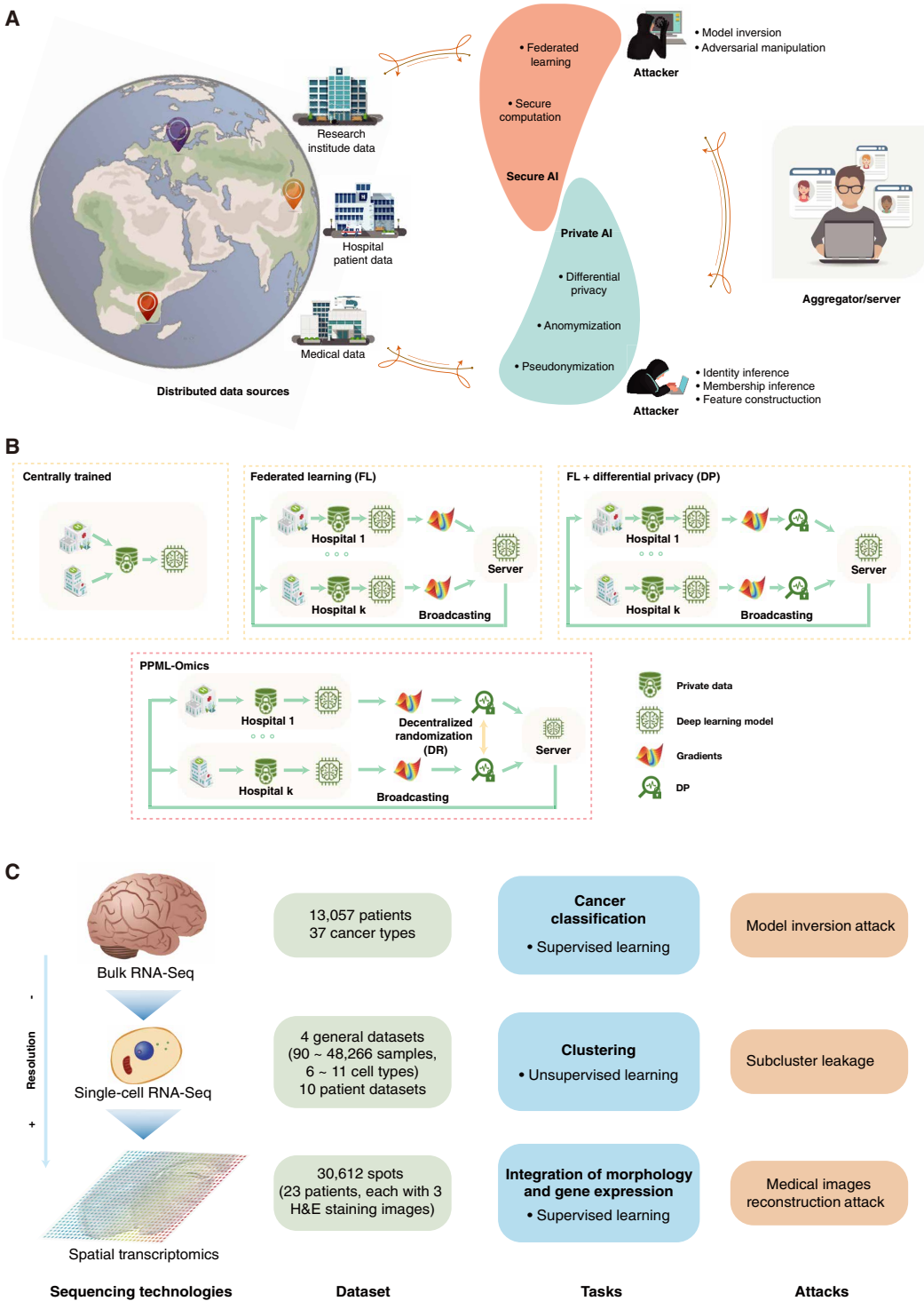

Fig. 1. PPML-Omics: A privacy-preserving federated machine learning method protects patients' privacy in omic data. (**A**) Schematic overview of the relationships and interactions between distributed data owners, aggregators, attackers, and techniques in the field of secure and private AI. (**B**) Schematic overview of different methods, including centrally trained method, federated learning (FL), FL with differential privacy (DP), and PPML-Omics. (**C**) Illustration of three representative tasks, datasets, and attacks of omic data here for demonstrating the utility and privacy-preserving capability of PPML-Omics, including the (i) cancer classification with bulk RNA-seq, (ii) clustering with scRNA-seq, and (iii) integration of morphology and gene expression with spatial transcriptomic.

(*32*). However, applying blockchain for deep learning with FL in practice is still a challenging field due to the high communication and computing costs (*33*). While cryptography-based deep learning techniques offer promising solutions to address privacy and security concerns in ML, they also come with certain drawbacks and challenges, such as communication overhead, key management, and security. Cryptographic systems rely on secure key management to protect the encryption keys. If the key management process is compromised, then the entire security of the system can be jeopardized (*34*).

Unlike cryptography-based methods, there is an alternative solution known as differential privacy (DP). It was introduced as a means to strike a balance between privacy and utility (section S1.2) (*35*). With DP, it becomes possible to safeguard privacy by sacrificing a certain amount of data utility, ensuring continuous protection even if the model is compromised by an attacker. Consequently, DP has been increasingly used, similar to cryptographic methods, to bolster privacy protection in the context of FL. The topic of FL with DP was explored extensively in the literature (*36*–*40*). However, most of the aforementioned articles only reported theoretical analyses of their framework or tested on classical datasets in the field of computer science, such as MNIST and CIFAR-10, whereas only a few of them applied their frameworks to real biological datasets that had more complex properties and greater intrinsic noise.

Among those works that applied FL, MPC, MHE, and DP to omic data analysis, most of them used either the cryptographic techniques (*41*–*43*) or the DP notion (*31*, *41*, *44*–*48*) to provide formal privacy guarantees for the participants in the research of single-nucleotide polymorphisms (SNPs), genome-wide association study (GWAS), and differential gene expression analysis (*49*), which are relatively narrow and specific problems in genomics studies, and whose data are obtained by postprocessing the raw sequencing data (section S1.3). In addition, the methods in those articles could only be shown to be applicable to statistical solutions or traditional ML solutions in GWAS. In addition to SNPs, raw sequencing data saved in the matrix format and generated by high-precision and quantitative multidimensional sequencing technologies contain much more sensitive information. Apart from that, only swarm learning (*32*) discussed the application of FL with blockchain technology in the analysis of omic data. But with the notation of DP, to our knowledge, only Islam *et al*. (*50*) discussed the application of DP in DL for breast cancer status and cancer type classification, and drug sensitivity prediction, and only one work discussed the application of FL-DP in cancer prediction as a solution to the competition hosted by iDASH (integrating Data for Analysis, Anonymization, SHaring) National Center for Biomedical Computing in 2020. Apart from that, there is no more work that has systematically studied and delved into the privacy protection of sequencing data from a bigger picture with the DP notation, even though raw sequencing data contains much more private information about patients than SNPs and GWAS. Meanwhile, the state-of-the-art work related to applying DP and MPC protocols in other biological tasks only reported privacy protection in medical imaging (*51*–*53*), which could be difficult to generalize to omic data analysis tasks because omic data have very different characteristics from imaging data.

To find a solution that is more applicable to practical scenarios of biological problems, we proposed a robust and powerful PPML-Omics method by designing a decentralized version of the differential private FL algorithm (see Materials and Methods) (Fig. 1B). In essence, the gradients of locally trained federated ML models are obfuscated through DP and decentralized randomization (DR) mechanisms before aggregating them at a single and non-trusted party. We applied PPML-Omics to analyze and protect privacy in real biological data from three representative omic data analysis tasks, which were solved with three different but representative deep learning models. We demonstrated how to address the privacy concern in the cancer classification from TCGA with bulk RNA-seq (*7*), clustering with scRNA-seq (*54*), and the integration of spatial gene expression and tumor morphology with spatial transcriptomics (*55*, *56*). In addition, we examined in depth the privacy breaches that existed in all three tasks through privacy attack experiments and demonstrated that patients' privacy could be protected through PPML-Omics as shown in Fig. 1C. In each of these applications, we showed that PPML-Omics was able to outperform methods of comparison, demonstrating the versatility of the method in simultaneously balancing the privacy-preserving capability and utility in omic data analysis. Last, we proved the privacy-preserving capability of PPML-Omics theoretically (section S1.4), suggesting the first mathematically guaranteed method with robust and generalizable empirical performance in the application of protecting patients' privacy in omic data.

In summary, our contribution provides the following innovations. We introduced the DP concept and systematically studied the privacy problem of multi-omics analysis in the form of three notable application scenarios in biology. We proposed PPML-Omics to achieve a better trade-off between model performance and privacy-preserving capabilities by designing a decentralized version of the differential private FL with the DR protocol based on the Fisher-Yates shuffle algorithm (*57*). In addition, we demonstrated the training of three representative deep learning models on three challenging tasks of omic data analysis using PPML-Omics and addressed the privacy concern in all three tasks, which may benefit following researchers by reminding the privacy issues in analyzing omic data. Besides, we conducted extensive experiments and showed that PPML-Omics was compatible with a wide range of omic data and biological tasks. In addition, we examined the computational performance of models trained with PPML-Omics against models trained centrally on the accumulated dataset and models trained with the FL method, FL-DP method, and FL-MHE method to demonstrate the strength of PPML-Omics under various scenarios typical in omic data analysis. Last, we assessed the theoretical and empirical privacy guarantees of PPML-Omics and provided examples of applying state-of-the-art attacks against the models in the application of protecting patients' privacy in omic data.

RESULTS

PPML-Omics and threat models

A confederation of N ($\geq$3) hospitals wishes to train three deep learning models for three tasks as shown in Fig. 1 (B and C). Since those hospitals had neither enough data themselves nor the expertise to train models on that data, they sought the support of a model developer to coordinate the training on a central server. PPML-Omics is built on this scenario to meet the needs of federal learning training and preserve the confidentiality of the local data from attacks. In the training phase of this method, each hospital has its own private patient data as the data owner. We suppose that participants trust each other (at least semi-honest) and do not actively undermine

the learning protocol. Each hospital trains a local model with its own private data and exchanges its gradients with another hospital randomly with the DR protocol before sending the updates to the server. Under this setting, each hospital is partially trusted by each other such that the gradients could be exchanged with a randomly generated partner in each epoch while the raw data cannot be accessed directly. In each epoch of training, after the DR mechanism, all hospitals may not hold their own original gradients, but rather gradients from a randomly paired hospital. Then, all hospitals upload their gradients to a server that is not trusted by the hospitals as the server usually requires strong communication power and is controlled and maintained by a third party. After integrating all gradients by the server, the server model is updated and sent to all hospitals to update all local models. Furthermore, individual participants excluding all clients are assumed as potential attackers to actively try to extract private information from other participants' data, the transmitted gradients during the training phase of the FL method, or the released pretrained model due to curiosity. The DP-based privacy enhancement technique was introduced in PPML-Omics to prevent such behavior. Specifically, it bounds the worst-case privacy loss for a single patient in the dataset and provides privacy guarantees to prevent model inversion/reconstruction attacks on federation participants or model owners during inference by adding noise to the gradients passed in the FL method. PPML-Omics implements DP and DR protocols to provide client-level privacy guarantees, further potentially protecting patient-level privacy. At the end of the training, all participants will hold a copy of the fully trained final model.

In Application 1, the participants try to train a model for cancer-type classification based on gene expressions from in-house bulk RNA-seq data. The server would release the final trained model and make it available to potential users. Since the released model may remember a large amount of gene expression information that is closely related to the cancer type from the patients. We assume that the attacker has auxiliary information, such as knowing that patient 1 is of cancer type A and has participated in the training of the released model. Thus, by performing the model inversion attack (MIA), the attacker could roughly know the gene expression of patient 1, resulting in a potential patient privacy breach.

In Application 2, the participants try to train a model for unsupervised clustering based on gene expressions from scRNA-seq data. The server would release the final trained model and make it available to potential users. We assume that in a real-world application scenario, users will have their own scRNA-seq data and also have our published model trained with PPML-Omics with different privacy budgets ϵ. The user's treating physician needs to analyze the user's clustering results from scRNA-seq data to determine the cell composition for medical judgment. However, the user does not want to present her clustering results to the treating physician with 100% accuracy due to privacy concerns. Since the user does not have extensive medical knowledge, the user cannot modify the clustering results by herself and can only hide some of the details of the clustering results by selecting different embedding models trained with PPML-Omics with different privacy budgets ϵ.

In Application 3, the participants try to train a model for predicting the spatial gene expression based on high-resolution images of hematoxylin and eosin (H&E) staining tissue from spatial transcriptomics data. The server would release the final trained model and make it available to potential users. Since the medical images used in the training phase contain lots of patients' privacy, we assume that an attacker will use the improved deep leakage from gradient (iDLG) method to reconstruct the images in the training dataset by stealing the gradient transmitted between the client and the server, resulting in a potential patient privacy breach.

Application 1. Cancer classification with bulk RNA-seq

Bulk RNA-seq can directly disclose patients' gene expression, while ML models trained with bulk RNA-seq can indirectly leak patients' signature-expressed genes (*48*). Here, we collected data from TCGA (see Materials and Methods and table S1) and compared computing resource requirements, privacy-preserving capabilities, and utility of data of different methods by working on the cancer classification task with gene expression as inputs.

To study the robustness and demonstrate the advantages of computational resources required by PPML-Omics, we used this task as a benchmark for the profiling analysis to test different methods (level 1: with or without FL, DP, and DR) against varied ML networks (level 2) (Fig. 2A). Depending on the differences in level 1, we tested five methods, including the centrally trained method, which was trained with one model on the entire dataset pooled on a single machine; the FL method, which was trained with separate models on the individual data owners' subsets of the dataset following the protocol of FL; the FL-MHE method, in which the MHE mechanism was integrated on the FL method; the FL-DP method, in which the DP mechanism was integrated on the FL method; and PPML-Omics, in which the DR protocol was designed to achieve a better trade-off between the model performance and the privacy-preserving capability.

Furthermore, for each method in level 1, variants based on the fully connected neural network (FCN) (Fig. 2A) were tested (level 2), including the use of different numbers of hidden layers (H1: 1 hidden layer, H3: 3 hidden layers), activation functions (R: ReLU, S: sigmoid), dropout layers (D0: without dropout layer, D05: with dropout layer and $P = 0.5$) and values of end-to-end ϵ_T, which can be used to further calculate the ϵ_l for each client in each epoch for the 37-class classification task, where the end-to-end privacy budget ϵ_T was a hyperparameter (see Materials and Methods) and a smaller ϵ_T means that we have stricter requirements for privacy protection, resulting in the need to add more noise and lower model performance. Data visualization through t-distributed stochastic neighbor embedding (t-SNE) showed that the use of gene expression could effectively distinguish between different cancer types (Fig. 2B). We trained all models to converge on the same dataset and measured the privacy-preserving capability (see Materials and Methods) and required computational resources, including time, random-access memory (RAM) and graphics processing unit (GPU) memory (section S2.2). Following Li's definition of the privacy loss in (*58*), we used Jensen-Shannon (JS) divergence and the P value of Kolmogorov-Smirnov (KS) test (P_{KS}) between two distributions to quantify privacy leakage, where a larger JS divergence and a smaller P value indicate a more severe privacy leakage. Compared to the centrally trained method, the FL method did not exhibit significant privacy-preserving power from the JS divergence perspective (JS divergence > 0.1 and similar to the centrally trained method). Compared to that, the integration of the DP (FL-DP method) substantially improved the privacy-preserving power (JS divergence $\leq$ 0.01) as shown in Fig. 2C. The JS divergence gradually decreased as the end-to-end privacy budget ϵ_T decreased (Fig. 2F), indicating that

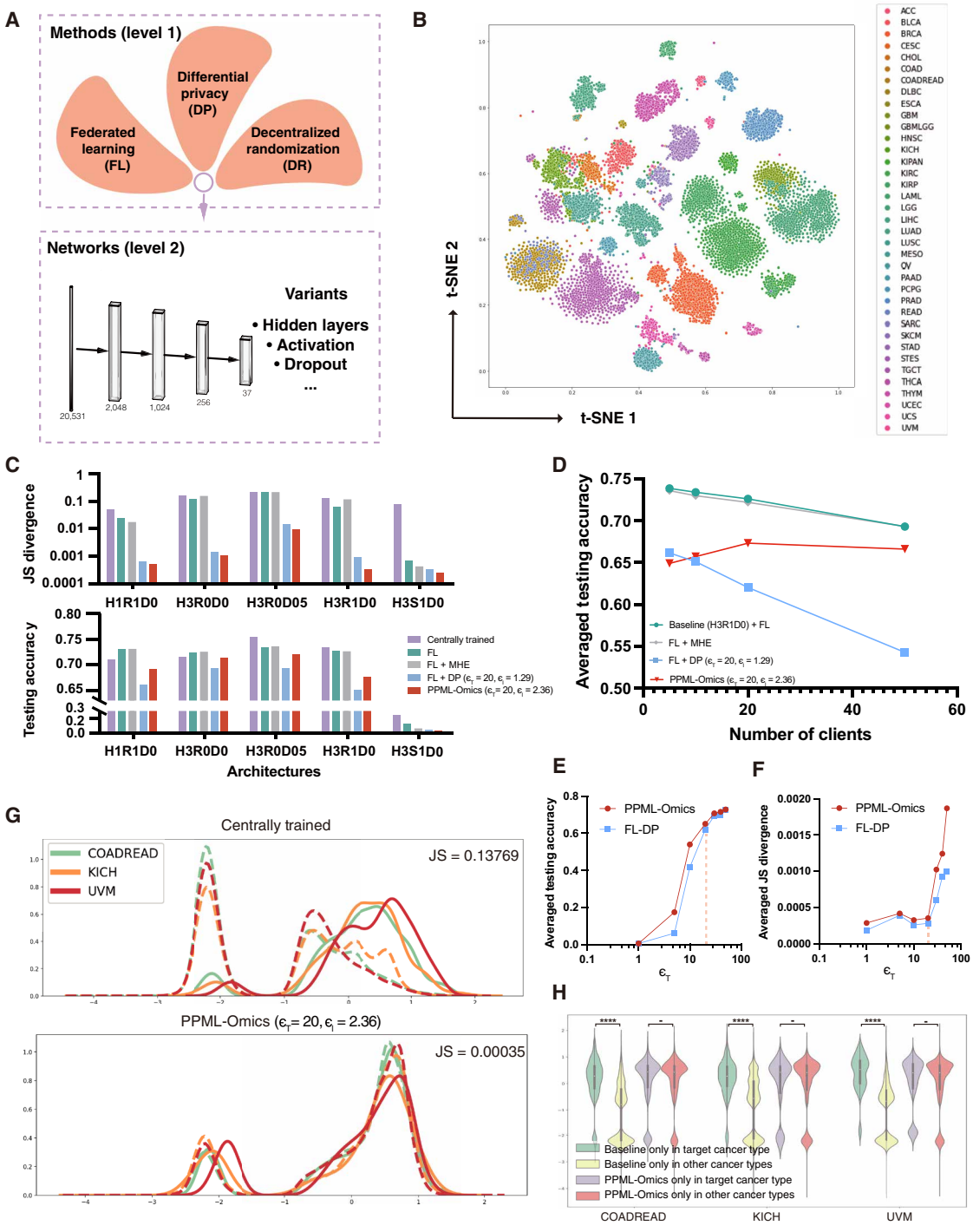

Fig. 2. Results of cancer classification with bulk RNA-seq in Application 1. (**A**) Illustration of the relationship between methods (level 1) and networks (level 2). (**B**) t-SNE plot on all patients' data from TCGA, each data point represents one patient and colors represent cancer types. (**C**) Profiling analysis of different methods, including the centrally trained method, the FL method, the FL-DP method, and PPML-Omics against different networks [level 2 in (A)] with varying numbers of hidden layers (H1: 1 hidden layer, H3: 3 hidden layers), activation function (R: ReLU, S: sigmoid), the dropout layer (D0: without dropout layer, D05: with dropout layer and $P = 0.5$). (**D**) The effect of the number of clients on application 1 of four methods. (**E**) The effect of the value of ϵ_T on the averaged testing accuracy of PPML-Omics and FL-DP. (**F**) The effect of the value of ϵ_T on the averaged Jensen-Shannon (JS) divergence of PPML-Omics and FL-DP. (**G**) The distribution of z-score normalized expression of reconstructed signature genes in the target cancer type (solid lines) and other cancer types (dashed lines) by the centrally trained method (baseline) and PPML-Omics from three representative cancers (COADREAD: colorectal adenocarcinoma, KICH: kidney chromophobe, UVM: uveal melanoma). (**H**) The violin plot of the distribution of z-score normalized expression of reconstructed signature genes in the target cancer type and other cancer types specifically identified by the centrally trained method and PPML-Omics. The test performed was a two-sided Kolmogorov-Smirnov test, and the P value annotation legend is the following: ****$P \leq 0.00001$, -$P > 0.001$. Exact P values are the following: Baseline only on COADREAD, $P = 5.98454 \times 10^{-59}$; PPML-Omics only on COADREAD, $P = 0.03325$; baseline only on KICH, $P = 3.59413 \times 10^{-34}$; PPML-Omics only on KICH, $P = 0.19957$; baseline only on UVM, $P = 2.14898 \times 10^{-131}$; PPML-Omics only on UVM, $P = 0.00144$.

adding higher noise (smaller ϵ_T) allowed for stronger privacy protection. As a trade-off between privacy protection and data utility, $\epsilon \leq 5$ is usually chosen in practice as in the handbook "Disclosure Avoidance for the 2020 Census: An Introduction" (see Materials and Methods) (*59*). However, there is no definite range of specific values for ϵ and the value of ϵ is dependent on the task and the data and needs to be selected based on a comprehensive consideration of the privacy-preserving requirement and the utility. In Application 1, $\epsilon_T = 20$ balanced the utility (Fig. 2E) and the privacy-preserving capability (Fig. 2F) of PPML-Omics. Therefore, with $\epsilon_T = 20$, PPML-Omics (JS divergence = 0.00035, $P_{KS} = 0.52341$) showed a clear advantage over the centrally trained model (JS divergence = 0.13769, $P_{KS} = 8.21642 \times 10^{-46}$), the FL method (JS divergence = 0.06169, $P_{KS} = 1.66231 \times 10^{-21}$), and the FL-MHE method (JS divergence = 0.10567, $P_{KS} = 2.01988 \times 10^{-39}$) when we need to meet both acceptable computational requirements (table S2) and privacy protection capabilities (Fig. 2C). Overall, PPML-Omics with $\epsilon_T = 20$ could meet the need for privacy protection under the practical scenario in Application 1 and the computational resource requirements of PPML-Omics were notably user-friendly.

Adding noise to gradients with DP has an amplification effect (*60*), meaning that the more times of adding noise, the greater the effect on model performance. Consequently, we need to handle the situation with multiple users when integrating DP into FL methods in real-life scenarios. To study the impact of the number of clients on the model performance with different methods and demonstrate the advantage of PPML-Omics, we tested the model performance by varying the number of clients on the same dataset. Since the integration of the DP mechanism requires adding noise to the gradient, increasing the number of clients had a notable impact on reducing the model performance of the FL-DP method (Fig. 2D). As a consequence, a large number of clients in real life would be the biggest challenge for applying the FL-DP method (*52*). However, the integration of the DR protocol could effectively weaken the performance degradation caused by increasing the number of clients, indicating that PPML-Omics could maintain better performance with a larger number of clients ($\geq$10), making it a feasible solution in practical applications.

Another obvious advantage of PPML-Omics was the better data utility, meaning that PPML-Omics could retain a higher level of data utility while protecting data privacy. To show the data utility of PPML-Omics in the cancer classification task, in the ablation study, we compared the averaged accuracy and macro-F1 score on the testing dataset of the centrally trained method, the FL method, the FL-DP method, the FL-MHE method, and PPML-Omics (Fig. 2C and tables S2 and S5). The FL method achieved slightly worse performance (accuracy = 0.72830) than the centrally trained method (accuracy = 0.73430) ($P = 0.18577$ for one-sided Student's t test). The FL-MHE method also achieved similarly good utility (accuracy = 0.73902) as the centrally trained method ($P = 0.76656$ for one-sided Student's t test). Meanwhile, the FL-DP method under $\epsilon_T = 20$ got the worst data utility (accuracy = 0.62136). With the same end-to-end privacy guarantee ($\epsilon_T = 20$), PPML-Omics showed significantly better utility (accuracy = 0.67713) than the FL-DP method (accuracy = 0.62136) ($P = 0.0006$ for one-sided Student's t test). In summary, PPML-Omics achieved good utility while preserving the privacy of gene expression.

Different cancer types have specifically expressed genes, which can be reconstructed by the attacker from the released models. Suppose that the attacker has auxiliary information, such as knowing that patient 1 is of cancer type A and has participated in the training of the released model. Thus, by performing the MIA, the attacker could roughly know the gene expression of patient 1, thus compromising the potential privacy of the patient as part of the model training data. To investigate the privacy protected by PPML-Omics and understand the biological meaning behind the JS divergence, we adopted the MIA for cancer classification models (see Materials and Methods). In other words, we tried to optimize a gene expression vector that could give the highest prediction probability for a particular cancer type on the target model by using MIA. Then, we analyzed the significantly expressed genes in the optimal gene expression vector. If the significantly expressed genes were cancer type-specific (with high expression in the target cancer type and low expression in other cancer types), then we could conclude that the target model leaked privacy. As shown in Fig. 2G, the most significant genes reconstructed from the centrally trained method with H3R1D0 showed significantly different distribution (JS divergence = 0.13769, $P_{KS} = 8.21642 \times 10^{-46}$) on the z-score normalized real expression between two groups (solid lines for the target cancer type and dashed lines for other cancer types) compared to the one with PPML-Omics ($\epsilon_T = 20$) (JS divergence = 0.00035, $P_{KS} = 0.52341$), suggesting that the MIA on the centrally trained method could reconstruct genes with significantly different expression levels (privacy leakage) on the target cancer type and other cancer types. In other words, based on the published centrally trained models, the attacker could reconstruct the corresponding specifically expressed genes for each cancer type. In contrast, it was impossible to accurately reconstruct the corresponding specifically expressed genes for each cancer type from the published models using PPML-Omics. A significant difference in the expression distribution of genes specifically reconstructed by the centrally trained method in the target cancer type and other cancer types could be observed (COAD-READ: $P_{KS} = 5.98454 \times 10^{-59}$, KICH: $P_{KS} = 3.59413 \times 10^{-34}$ and UVM: $P_{KS} = 2.14898 \times 10^{-131}$) (Fig. 2H), implying that genes reconstructed for each cancer type tended to have a higher expression in the corresponding cancer type. In contrast, the expression distribution of genes reconstructed by attacking PPML-Omics was very similar in the target cancer type and in other cancer types (COAD-READ: $P_{KS} = 0.03325$, KICH: $P_{KS} = 0.19957$ and UVM: $P_{KS} = 0.00144$) (Fig. 2H), implying that the genes reconstructed for each cancer type were not strongly correlated with the cancer type in terms of expression. A similar observation could be found across all cancer types that reconstructed genes by attacking the model trained with the centrally trained method showed significantly different expression levels between the target cancer type and other cancer types whereas reconstructed genes by attacking the model trained with PPML-Omics did not show such a significant difference (section S2.3, figs. S1 to S3, and tables S3 and S4). Overall, we showed that PPML-Omics protects genomic privacy by obfuscating sensitive gene identification.

Cryptography-based FL method could protect the security during the training phase, but this could also mean no protection at all, especially when the key leaks. We implemented an FL method based on MHE, in which models and parameters were encrypted during the communication, training, and inference phases using the homomorphic encryption (HE) method based on the Cheon-Kim-Kim-Song (CKKS) cryptographic scheme (see Materials and Methods) (*61*). The clients and server have public and private keys for secure

encryption and decryption. As shown in tables S2 and S5, the model trained with the FL-MHE method achieved similar accuracy (accuracy = 0.73902) as the centrally trained method. Under the scenario of Application 1, the model could not be decrypted if assuming that the attacker could not obtain the public and private keys, thus the security and privacy are completely protected. Once the attacker obtained the public and private keys, the server model could be fully decrypted and the privacy leakage (JS divergence = 0.10567, $P_{KS} = 2.01988 \times 10^{-39}$) was as serious as that of the centrally trained method as shown in Fig. 2C, fig. S2, and table S5 even though the encryption and decryption processes introduced approximation errors and protected very little privacy. Meanwhile, the FL-MHE method required additional computational resources due to the encryption and decryption as shown in table S2. Overall, PPML-Omics showed a clear advantage over the FL-MHE method under our scenario in Application 1.

Application 2. Clustering with scRNA-seq data

scRNA-seq is a revolutionary technology to quantify gene expression in thousands and even millions of cells in parallel, which is powerful in identifying cell populations and dissecting the composition of each population in biological samples (*62*). Applying scRNA-seq on primary tumors is not only able to decipher tumor cell heterogeneity (*63*) but can also uncover the specificity of tumor microenvironment (*64*) in each patient, which may lead to a personalized treatment strategy (*65*). For instance, immune infiltration in the tumor which could be determined by scRNA-seq is a good response indicator to immune blockade-based therapy. Thus, information on cell populations is critical in protecting patients' privacy when analyzing scRNA-seq data. Using the extracted features rather than the direct gene expression may avoid the leak of the specifically expressed gene while still harboring the information for each cell population in the patients. Even so, obtaining accurate cell population results on patients' tissue could also violate privacy. Here, we used the low-dimensional features extracted from gene expression vectors of scRNA-seq data by Auto-encoder as input for K-means clustering (Fig. 3A). We then applied PPML-Omics to see how it protected patients' privacy from the results of clustering, by evaluating the data utility and the privacy-preserving capability in terms of cell type classification and composition quantification based on clustering.

The number and composition of cells included in the scRNA-seq data vary greatly depending on the sample. To assess the robustness of PPML-Omics for different sample sizes in the clustering task, we evaluated it on three benchmark datasets (Yan, Pollen, and Hrvatin) with varying cell numbers, ranging from 90 to 48,266 (table S8). Following Tran's work (*66*), we compared the adjusted Rand index (ARI), normalized mutual information (NMI), cluster accuracy (CA), and Jaccard index (JI) (see Materials and Methods) on the three datasets with different methods (Table 1), including the centrally trained method, the FL method, the FL-DP method, the FL-MHE method, and PPML-Omics, where a larger value means a better clustering result. For all three datasets, PPML-Omics achieved promising performance compared to the FL-DP method under all four evaluation metrics (Yan: $P = 5.23 \times 10^{-3}$, Pollen: $P = 9.12 \times 10^{-4}$, and Hrvatin: $P = 1.01 \times 10^{-5}$ from one-sided Student's t test) with the same privacy budget ($\epsilon_T = 5$). Also, the difference between the clustering result from PPML-Omics and that from the centrally trained model is relatively minor (Yan: $P = 0.048$, Pollen: $P = 0.194$, and Hrvatin: $P = 0.012$ from one-sided Student's t test), indicating that our framework even had the potential to approach the centrally trained method in clustering task. To ensure that the performance of PPML-Omics was at the same level as those commonly acknowledged tools in the clustering task with scRNA-seq data, we compared PPML-Omics with the existing state-of-the-art tools, including Seurat (*67*), SC3 (*68*), CIDR (*69*), and SINCERA (*70*) (sections S3.2 and S3.3, fig. S4, and table S7). PPML-Omics also achieved competitive utility, proving that our method achieved similar performance as the most commonly used tools for the clustering task. To further investigate the effect of FL, DP, and DR protocols on the clustering task, we conducted an ablation study (Fig. 3B) [centrally trained method at the first column, FL method ($\epsilon_T = 5$, $\epsilon_l = 0.23$) at the second column, FL-DP method ($\epsilon_T = 5$, $\epsilon_l = 0.33$) at the third column, and PPML-Omics at the fourth column] on these three datasets of different sizes and visualized the clustering results with Uniform Manifold Approximation and Projection (UMAP) algorithm that projected the internal representations into a two-dimensional space. In conclusion, PPML-Omics could qualitatively and quantitatively compete or even outperform other methods in terms of utility.

When we obtain the scRNA-seq data of patients, we can learn the composition of the cells based on the clustering results. In extreme cases, suppose the clustering method is good enough and can 100% correctly distinguish different types of cells into clusters, it might potentially violate the patient's privacy by leaking some sensitive sub-cell types. Therefore, it is necessary to reasonably adjust the resolution of the clustering method to properly hide some small clusters (sub-cell types) from the final clustering results to achieve privacy protection according to the prior privacy requirements. In other words, for a method, if we cannot observe some sensitive small clusters (sub-cell types) in the final clustering results, then we could conclude that the method protects the patient's privacy. In addition, if the clustering results for major cell types are reasonable, then we could conclude that the method preserves an acceptable degree of data utility while protecting privacy. To study the privacy-preserving power of PPML-Omics, we analyzed a public scRNA-seq dataset with 43,817 cells from 10 patients with colon cancer (*71*), in which the cells were classified into five major cell types (hB: B cells, hI: innate lymphoid cells, hM: myeloid cells, hT: $CD4^+$ T cells, and $CD8^+$ T cells) and 38 sub-types (section S3.5 and tables S9 and S10). We applied three methods for clustering, namely, the centrally trained method and two of PPML-Omics with different noise levels ($\epsilon_T = 10$, $\epsilon_l = 0.25$) and ($\epsilon_T = 5$, $\epsilon_l = 0.12$). With two patients (P0408 and P0410) as examples, the major cell types were successfully identified with all three methods (Fig. 3C), indicating that the overall clustering results of the three methods were reasonable, thus preserving the utility of the data. However, the clustering results from the three methods showed notable differences in the resolution of the sub-cell types and we could use it as an evaluation of the privacy-preserving capability similar to the JS divergence in Application 1. Comparing both our methods ($\epsilon_T = 10$, $\epsilon_l = 0.25$) and ($\epsilon_T = 5$, $\epsilon_l = 0.12$) with the centrally trained method, four subclusters (sub-cell types) were undetected for P0408 and 6 subclusters (sub-cell types) were undetected for P0410 (tables S9 and S10), suggesting that the local differences between subclusters (sub-cell types) could be diluted due to the noise added in PPML-Omics and then those sub-cell types could be integrated into major clusters, thus achieving privacy protection as a consequence.

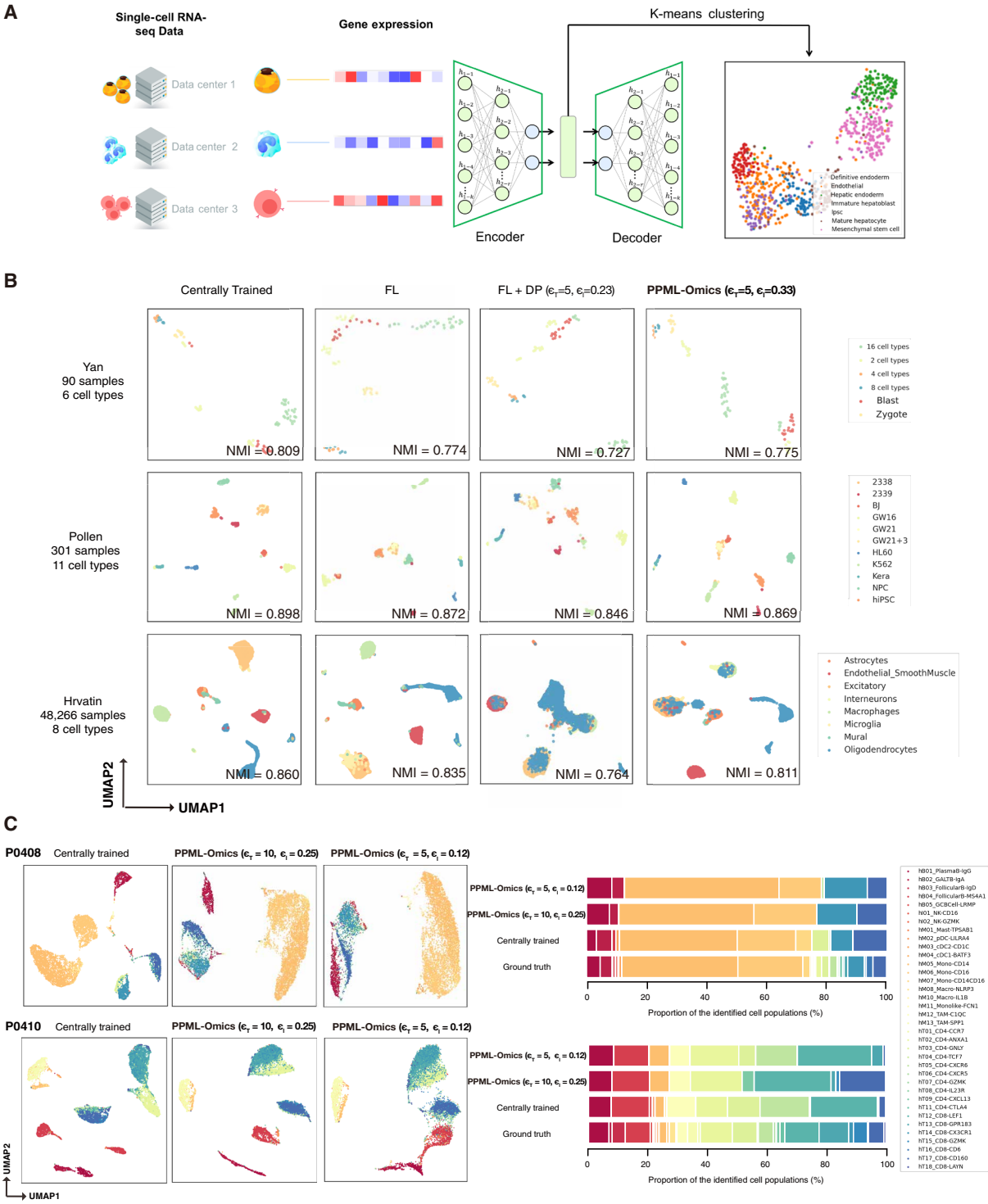

Fig. 3. Results of clustering with scRNA-seq in Application 2. (**A**) Architecture of the backbone (Auto-encoder). Gene expression vectors were fed into the Auto-encoder for feature reduction and selection, after which the low dimensional features were used for the K-means clustering. (**B**) The clustering results visualized by the Uniform Manifold Approximation and Projection (UMAP) of three datasets with an increasing number of samples (Yan with 90 samples from 6 cell types, Pollen with 301 samples from 11 cell types, and Hrvatin with 48,266 samples from 8 cell types) generated by four methods, including the centrally trained method, the FL method, the FL-DP method, and PPML-Omics. For these three datasets, all clustering results given by PPML-Omics showed a similar visual pattern to the centrally trained method, indicating the perfect utility of PPML-Omics. (**C**) The clustering results visualized by the UMAP of two patients (P0408 and P0410) and the proportion plot of major clusters and subclusters generated by different methods, including the centrally trained method, and PPML-Omics with ($\epsilon_T = 10$, $\epsilon_I = 0.25$) and ($\epsilon_T = 5$, $\epsilon_I = 0.12$), indicating that PPML-Omics could protect patients' privacy by removing the local information of subclusters.

Table 1. Comparison of clustering result on multiple scRNA-seq datasets.

Dataset	Method	ARI	NMI	CA	JI
Yan (90 samples 6 cell types)	Baseline (centrally trained)	0.707	0.809	0.777	0.616
	FL	0.644	0.774	0.724	0.555
	FL + MHE	0.617	0.812	0.711	0.526
	FL + DP ($\epsilon_T = 5, \epsilon_l = 0.23$)	0.560	0.727	0.689	0.479
	PPML-Omics ($\epsilon_T = 5$, $\epsilon_l = 0.33$)	0.636	0.775	0.722	0.549
Pollen (301 samples 11 cell types)	Baseline (centrally trained)	0.847	0.898	0.883	0.759
	FL	0.785	0.872	0.824	0.673
	FL + MHE	0.731	0.850	0.766	0.611
	FL + DP ($\epsilon_T = 5$, $\epsilon_l = 0.23$)	0.748	0.846	0.801	0.632
	PPML-Omics ($\epsilon_T = 5$, $\epsilon_l = 0.33$)	0.802	0.869	0.836	0.699
Hrvatin (48,266 samples 8 cell types)	Baseline (centrally trained)	0.816	0.860	0.851	0.741
	FL	0.742	0.835	0.775	0.649
	FL + MHE	0.805	0.815	0.831	0.731
	FL + DP ($\epsilon_T = 5$, $\epsilon_l = 0.23$)	0.737	0.764	0.790	0.645
	PPML-Omics ($\epsilon_T = 5$, $\epsilon_l = 0.33$)	0.737	0.811	0.784	0.645

Using a larger end-to-end ϵ_T means better utility and greater resolution on the clustering results (more sensitive subclusters could be observed) while using a smaller ϵ_T means that we could sacrifice a certain level of utility to protect sensitive subclusters in the clustering results. For potential users of PPML-Omics, the value of ϵ_T should be adjusted according to the needs of the actual application scenario to achieve both good utility and privacy protection as shown in (fig. S8).

Application 3. Integration of tumor morphology and gene expression with spatial transcriptomics

Sequencing data in Applications 1 and 2 could reveal much sensitive information about patients, and the potential risk of privacy disclosure is also very high for spatial transcriptomics as one of the highest-resolution sequencing technologies. In addition, spatial transcriptomics requires the usage of tissue images, which adds additional privacy leakage to some extent. ST-Net (*56*) is an ML model for predicting local gene expression from H&E-stained pathology slides trained on a dataset of 30,612 spatially resolved gene expression data matched to histopathology images from 23 patients with breast cancer. Unlike the classical task of spatial transcriptomics, ST-Net predicts the spatially resolved transcriptome of tissue directly from tissue images, while the gene expression measured from spatial transcriptomics is used as the ground truth in the training phase. Here, to show that PPML-Omics realized competitive utility and to investigate how PPML-Omics protects privacy in those histopathology images on this task, we applied PPML-Omics to integrate tumor morphology and gene expression with spatial transcriptomics by incorporating the ST-Net model as the backbone network into PPML-Omics and compared with Hist2ST (*72*) and Hist2Gene as shown in Fig. 4A and Table 2.

To demonstrate that ST-Net under PPML-Omics has the same level of utility compared to pure ST-Net in predicting local gene expression from pathology slides, we visualized the prediction results and quantitatively compared the similarity between the predicted results and the ground truth by calculating the mean square error (MSE) as an evaluation metric, where a larger MSE means a worse prediction. From the visualization as shown in Fig. 4B, both PPML-Omics (at the sixth row) and the centrally trained method (at the fourth row) obtained good results from a visual perspective compared to the ground truth (at the third row), effectively predicting the local expression of *FASN* and other cancer marker genes, including *HSP90AB1* and *PABPC1* (section S4.2 and figs. S6 and S7) with the histopathology images of five patients (BT23269 C1, BT23277 E1, BT23377 C1, BT23901 C2, and BT23944 E1) as inputs, which could be supported by the sequencing data from spatial transcriptomics (ground truth). Furthermore, the local gene expressions predicted by PPML-Omics and the centrally trained method were both aligned well to the tumor region (black) and normal region (white) annotated in the annotation (at the second row), indicating that the

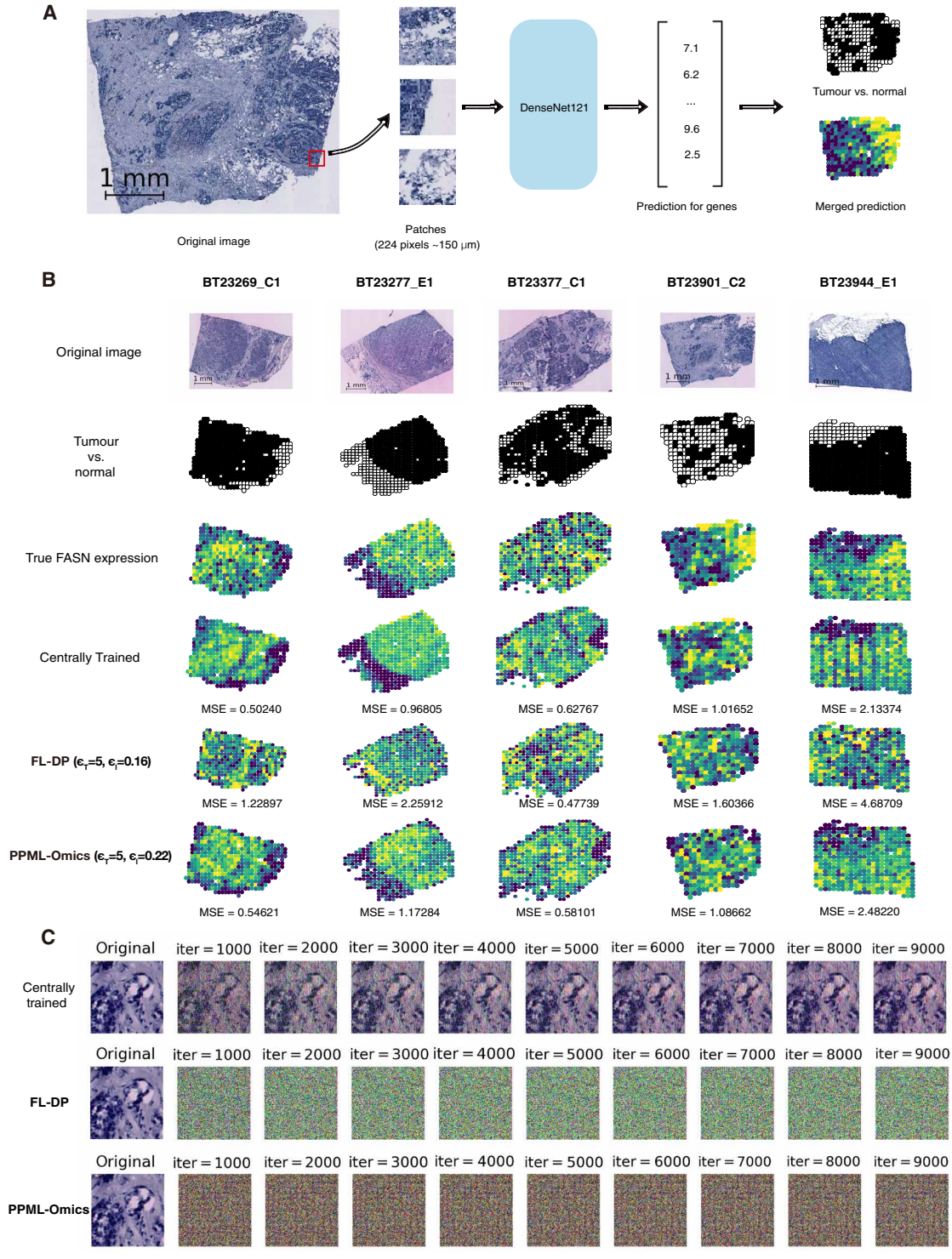

Fig. 4. Results of integration of tumor morphology and gene expression with spatial transcriptomics in Application 3. (**A**) Pipeline for predicting local gene expression from high-resolution tissue image referred by ST-Net. The patches (224 × 224) extracted from the original image were fed into the DenseNet121 and the local expression of selected cancer marker genes was predicted and compared with the real local expression (ground truth) from the spatial transcriptomics data. (**B**) The results of five samples (BT23269 C1, BT23277 E1, BT23377 C1, BT23901 C2, and BT23944 E1) showed the original image, the binary labels of the tumor (black) and normal regions (white), the real expression of FASN, and predicted results by the centrally trained method, FL-DP ($\epsilon_T = 5$, $\epsilon_I = 0.16$), and PPML-Omics ($\epsilon_T = 5$, $\epsilon_I = 0.22$). All predictions show a similar visual pattern with the ground truth and both the centrally trained method and PPML-Omics give similar good performance in terms of MSE. (**C**) Image reconstruction attack with improved iDLGs on the centrally trained method, FL-DP, and PPML-Omics.

Table 2. Comparison of different methods in the integration task of spatial gene expression and tumor morphology with spatial transcriptomics.

Method	Averaged testing MSE
Hist2ST, centrally trained	0.91053
Hist2Gene, centrally trained	0.88285
Baseline (ST-Net, centrally trained)	0.90904
Baseline + FL	0.93120
Baseline + FL + MHE	0.90904
Baseline + FL + DP ($\epsilon_T = 0.1$, $\epsilon_I = 0.003$)	NA
Baseline + FL + DP ($\epsilon_T = 0.5$, $\epsilon_I = 0.016$)	NA
Baseline + FL + DP ($\epsilon_T = 1$, $\epsilon_I = 0.032$)	1557.54465
Baseline + FL + DP ($\epsilon_T = 5$, $\epsilon_I = 0.160$)	1.29931
PPML-Omics ($\epsilon_T = 0.1$, $\epsilon_I = 0.004$)	3.22789
PPML-Omics ($\epsilon_T = 0.5$, $\epsilon_I = 0.022$)	1.08788
PPML-Omics ($\epsilon_T = 1$, $\epsilon_I = 0.045$)	0.98808
PPML-Omics ($\epsilon_T = 5$, $\epsilon_I = 0.225$)	0.93005

ST-Net integrated with PPML-Omics still had the power to predict the local gene expression accurately from tissue images. In addition, with the same small privacy budget $\epsilon_T = 1$ and 5, the prediction of PPML-Omics was significantly more accurate (MSE = 0.98808 at $\epsilon_T = 1$ and MSE = 0.93005 at $\epsilon_T = 5$) than the FL-DP method (MSE = 1557.54465 at $\epsilon_T = 1$ and MSE = 1.29931 at $\epsilon_T = 5$) with $P = 8.06039 \times 10^{-29}$ and $P = 3.10063 \times 10^{-14}$ for one-sided Student's t tests respectively and achieved close performance of the centrally trained method (MSE = 0.90904) (Table 2 and table S13). Thus, PPML-Omics achieved good utility in Application 3.

It is well known that medical imaging data contains much potentially sensitive information (*11*), such as tissue patterns and lesions, which could compromise patients' privacy. Several studies have shown that ML models trained based on medical images remember much information about the images in the training dataset, and attackers could perform image reconstruction attacks with the published ML models to obtain the original training images (*73*–*75*). To see whether PPML-Omics protects privacy in the histopathology images compared to the centrally trained method, we used the iDLG (*76*, *77*) (see Materials and Methods) to simulate an attacker reconstructing the training images by stealing the gradients passed between machines under the FL framework (approximating a centrally trained method when the number of clients is 1). In each attack, we initialized a noisy input (dummy data), used the local gene expression as the ground truth, computed the prediction of the model on the noisy input, and calculated the gradient. Then, we updated the noisy input with the gradient. In other words, same as in Application 1, we were trying to optimize an input image that could get the most similar prediction results to the ground truth. We performed the iDLG attack separately for the centrally trained method and PPML-Omics and showed the input image every 1000 iterations, as shown in Fig. 4C. Notably, in terms of visual results, the iDLG attack could effectively reconstruct the training images with the centrally trained method, while in PPML-Omics, the iDLG attack was effectively blocked due to the addition of noise to the gradients with the DP mechanism. Overall, PPML-Omics protected privacy by blocking the reconstruction of sensitive histopathology images.

Theoretical proof of the privacy-preserving power of PPML-Omics

With the three applications above, we have empirically demonstrated our method's privacy-preserving capability and utility. We further theoretically proved the privacy-preserving capability of PPML-Omics with the DP notation. The central DP (CDP) model and the local DP (LDP) model are two commonly acknowledged models with the notation of DP. In the CDP model, a server trusted by users collects users' raw data (e.g., local updates) and executes a private mechanism for deferentially private outputs. The privacy-preserving goal is to achieve indistinguishability for any outputs w.r.t. two neighboring datasets that differ by replacing one user's data. The definition of DP (definition 4 in section S1.2) requires that the contribution of an individual to a dataset has not much effect on what the adversary sees. Compared to CDP, LDP has a stronger notion of privacy. In LDP, each user's data are required to be perturbed to protect privacy before collection. The CDP model assumes the availability of a trusted analyzer to collect raw data, while the LDP model does not rely on any trusted party because users send randomized data to the server. The CDP model protects the calculation result in the analyzer, so users need to trust the central server and send raw data to the server, which allows greater accuracy but requires a trusted analyzer which is impractical in most real cases. In the LDP model, we protect the data information in the single local device, so users only need to trust their single device and randomize local data before sending them to the analyzer. Although a trusted central analyzer is not required, the utility of the method is limited because we do lots of randomness on local data. Therefore, the advantage of using the DR protocol (definition 6 in section S1.4) is that we could balance the strength in both CDP and LDP, i.e., good performance of accuracy in the CDP model and strong privacy in the LDP model without relying on any trusted central party.

Under the FL, each client trains its model locally and sends the update in the form of gradients to a central server that would aggregate those updates into the central model. After updating the central model, the central server broadcasts the new model weight to all clients for updating all clients (Algorithm 1). With the

integration of DP in the FL framework, we manually chose several end-to-end ϵ_T and set δ as discussed in Materials and Methods. Given ϵ_T and δ_T, we applied the analytic Gaussian mechanism to calculate the optimal σ for perturbation (*78*) as in Algorithm 2. In other words, we could calculate the amount of noise that needs to be added based on the given ϵ_T and δ_l. The privacy guarantee of the analytic Gaussian mechanism was given in (theorem 1 in section S1.4). Thus, we applied the analytic Gaussian mechanism on a single client with the privacy guarantee proved to realize the DP mechanism.

Based on the postprocessing property (lemma 1 in section S1.4), the protocol P (defined in definition 6 in section S1.4) achieves the same privacy level as M (defined in lemma 1 in section S1.4) because A is executed by an untrusted analyzer, without protecting users' privacy. We wanted to obtain $M = S \circ R^n$ that could satisfy (ϵ_c, δ_c)-DP, meaning that we achieved the same privacy guarantee as the CDP. Thus, we focused on analyzing the indistinguishability for $M(X)$ and $M(X')$ where $X$and X' differ in one client's local vector such that we achieved privacy-preserving on the client level. Erlingsson *et al.* (*79*) proved that the privacy of M could be amplified. In other words, when each user applies the local privacy budget ϵ_l in R, M can achieve stronger privacy of (ϵ_c, δ_c)-DP with $\epsilon_c < \epsilon_l$. Hence, the DR protocol has a larger privacy budget for a local single client and needs less noise to achieve the same privacy model compared with the LDP model. In practice, we defined a privacy budget ϵ_c in a single epoch, and then we amplified this budget to the local client model with the local privacy budget ϵ_l.

Given target privacy parameters $0 < \epsilon_T < 1$ and $0 < \delta < 1$, to ensure (ϵ_T, δ_T)-DP over T mechanisms, it suffices that each mechanism is (ϵ_c, δ_c)-DP, where $\epsilon_c = \frac{\epsilon_T}{2\sqrt{\left\{2Tln\left(\frac{2}{\delta_T}\right)\right\}}}$ and $\delta_c = \frac{\delta_T}{2T}$. From the advanced composition theorem and the corollary (section S1.4), we can guarantee that Algorithm 2 satisfies (ϵ_T, δ_T)-DP after T epochs.

DISCUSSION

Overall, PPML-Omics is universal, meaning that PPML-Omics could be integrated with any ML model and applied to various biological problems. Thus, we applied PPML-Omics to three different ML models from simple to complex, including FCN, Anto-encoder, and DenseNet-121. Applying PPML-Omics to more complex deep learning models means that more computational resources are required to add noise, which may introduce more uncertainty in the performance of the models. Hence, whether PPML-Omics can be applied to more complex ML models and used to protect privacy in more complex data requires further research in the future. Furthermore, in addition to the three distinct resolutions of sequencing data analysis tasks discussed in this work, PPML-Omics holds the potential for application in other relevant fields, such as the integration of eQTL data in future endeavors.

Batch effects are a prevalent issue when dealing with multi-institutional omics data and necessitate careful consideration during data analysis. In traditional scenarios, omics data from various institutions are combined, and batch effects are mitigated using established methods before proceeding with the analysis. However, in the context of FL, where data remains distributed across multiple participants or centers, batch effects pose a distinct challenge that requires tailored solutions to ensure the robustness of the trained models. FL provides a framework for addressing batch effects, both through supervised and unsupervised learning methods. When these methods are extended to FL, they converge on a common objective: To identify and emphasize shared biological features across different batches, thereby facilitating the removal of batch effects and enhancing the model's predictive capabilities. Furthermore, note that batch effects are also a well-recognized challenge within the FL domain, and several solutions have been developed. Approaches such as FedBN in (*80*) and in (*81*) have been proposed to mitigate batch effects. In the case of PPML-Omics, which was developed on FL, it can also effectively address batch effects by incorporating these various solutions and methodologies by default (fig. S11). This ensures that PPML-Omics is equipped to handle batch effects in a decentralized and collaborative data analysis setting.

The integration of the DP mechanism requires additional computational resources, such as more GPU memory consumption and computation time, for calculating the magnitude of the noise and the gradient calculation with noise. However, we demonstrated in our simulated experiments that the computational burden imposed by PPML-Omics was insignificant compared to centrally trained methods, implying that potential users could easily deploy PPML-Omics for privacy-preserving omic data analysis without upgrading their existing computational devices.

It is commonly acknowledged that privacy protection and utility are two conflicting objectives. Thus, it is challenging to find a balance that allows the model to protect privacy without overly damaging the usability of the data. Therefore, PPML-Omics balances the privacy-preserving capabilities with utility as much as possible while also leaving some of the decision-making to our potential users. Depending on the user's actual needs, the user can adjust the end-to-end privacy budget (ϵ_T) in the method during the training phase or select the released model trained with PPML-Omics under different ϵ_T to achieve different levels of privacy protection. We acknowledge the fact that PPML-Omics cannot automatically select ϵ_T for potential users for specific biological problems but requires the user to manually select ϵ_T is a major weakness of PPML-Omics. However, there are potential reasons for this as different biological data have different characteristics as well as various inherent noise, and different deep learning models could tolerate different levels of noise. For example, for the three applications, we selected different values of ϵ_T to achieve appropriate privacy protection (e.g., $\epsilon_T = 20$ in Application 1, $\epsilon_T = 5$ in Application 2, and $\epsilon_T = 5$ in Application 3). It may take more time for users to try different values of ϵ_T to choose an appropriate one that fits the privacy requirement the most when applying PPML-Omics in practice. We also provided the relationship between various evaluation metrics and the value of ϵ_T in all three applications as a reference (Fig. 2, E and F, and figs. S8 and S9).

In summary, we have proposed a secure and PPML method by designing a decentralized version of the differential private FL algorithm (PPML-Omics). Besides, we have applied this method to analyze data from three representative omic data analysis tasks, which are solved with three different deep learning models, revisited and addressed the privacy concern in the cancer classification from TCGA with bulk RNA-seq, clustering with scRNA-seq, and the integration of spatial gene expression and tumor morphology with spatial transcriptomics. Moreover, we examined in depth the privacy

breaches that existed in all three tasks through privacy attack experiments and demonstrated that patients' privacy could be protected by PPML-Omics. In addition, we proved the privacy-preserving capability of PPML-Omics theoretically, suggesting the first mathematically guaranteed method with robust and generalizable empirical performance in protecting patients' privacy in omic data. We believe that PPML-Omics will attract the attention of future researchers on biological issues regarding data privacy and also try to protect data privacy by applying PPML-Omics in research. Our method's modularized and extendable nature has great potential to be developed collaboratively for different biological tasks and deep learning models and will shed light on the application and development of deep learning techniques in the privacy protection of biological data.

MATERIALS AND METHODS

Dataset preparation

For Application 1, we used TCGA dataset, accessible through the Genomic Data Commons Data Portal, and built our methods across 37 cancers with a total number of 13,057 patients. Each patient had the expression vector of 20,531 genes, and each expression vector was rescaled in logarithm with 10 according to the following equation, to speed up learning and lead to faster convergence. Also, the normalization procedure ensured that the gene expression contributing to the model was on an equal scale. Patients were randomly divided into a training dataset (80%) and a test dataset (20%) and the random split was repeated five times (section S2.1).

$$X_{\text{input}} = \log_{10}(X_{\text{original}})$$

For Application 2, to assess the representation from our privacy-preserved model on the scRNA-seq data, we tested our model on three published datasets with different sample sizes, all of which have expert-annotated labels from single-cell Decomposition using Hierarchical Autoencoder (scDHA) (*66*). The Yan (*82*) dataset refers to the human preimplantation embryos and embryonic stem cells. In this dataset, 90 cells were sequenced with the Tang protocol. We first log-transformed the gene expression values selected for the highly variable genes. Last, we scaled the dataset to unit variance and zero mean. The Pollen (*83*) dataset was sequenced with the SMARTer protocol. It contains 301 cells in the developing cerebral cortex from 11 populations. We downloaded it from the Hemberg Group's website (https://hemberg-lab.github.io/scRNA.seq.datasets/human/tissues/) and removed the low-quality data and cells with more mitochondrial genes and spike-in RNA. Then, we selected highly variable genes after logarithmic transformation. The Hrvatin (*84*) dataset contains 48,266 cells from 6- to 8-week-old mice, which were sequenced by DropSeq. After filtering the low-quality data, we performed the aforementioned logarithmic transformation, normalization per cell count, and highly variable gene selection steps. To further validate the privacy-preserving ability of PPML-Omics, we selected the scRNA-seq data on immune and stromal populations from patients with colorectal cancer (*71*). We visualized the global and local clusters regarding their specific macrophage and conventional dendritic cell subsets. The single-cell data were selected from patients with informed mechanisms of myeloid-targeted therapies in colon cancer, which can be found in "Data and materials availability" in Acknowledgements. After quality control, each patient has around 5000 cells and corresponding 13,538 genes (section S3.1).

For Application 3, to test the capability of our method in spatial transcriptomics, we used a spatial transcriptomics dataset of 30,612 spots, including 23 patients with breast cancer with four subtypes: luminal A, luminal B, triple-negative, and human epidermal growth factor receptor-2 (HER2)–positive (*56*). There were three high-resolution images of H&E staining tissue and their corresponding gene expression data for each patient. The number of spots for each replicate ranged from 256 to 712 depending on the tissue size and the diameter of a single spot was 100 μm arranged in a grid with 200 μm as the center-to-center distance. There were 26,949 mRNA species detected across the dataset and each spot was represented as a 26,949-dimensional vector containing the elements denoted by the number of gene counts (section S4.1).

Deep learning model design and training

All codes were implemented using the PyTorch library (*85*) and processed on one machine with 2 NVIDIA V100 GPUs and 252G RAM. All experiments were repeated on the pre-split training and testing batches until fully converged on the testing dataset, and all reported performance was averaged over five random splits.

For Application 1, we applied an FCN with three hidden layers as our benchmark ML network for PPML-Omics. The gene expression vector was fed into the network, and a ReLU layer was applied after each hidden layer to provide nonlinearity. Last, a soft-max layer was attached to the last hidden layer to get the final prediction. With PPML-Omics, the network was trained in a federated, secured, and privacy-preserving procedure to guarantee utility while preserving privacy at the same time. The multi-class cross-entropy loss defined the loss function

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic})$$

where M represents the number of cancers (classes), which is 37 for this task. Correspondingly, $y_{ic} \in \{0, 1\}$ represents the classification result for the ith sample, with 1 for a positive prediction, and 0 for a negative prediction. p_{ic} stands for the probability of the ith sample to be in class c.

For Application 2, the unsupervised framework for scRNA-seq clustering is composed of two steps. We first trained the Auto-encoder with the fixed number of gene expression value y as the input and outputted $\hat{y}$ with the same size as the input. We denoted d as the dimension of each patient's gene expression vector and N as the number of clients. We then optimized the Auto-encoder with the Mean squared error (MSE) loss

$$L(y,\hat{y}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{d}(y-\hat{y})^2$$

Then, we extracted the central feature vector and clustered it with K-means clustering.

For Application 3, we used ST-Net (*56*) as our baseline deep learning network and integrated it into PPML-Omics. In the ST-Net, we used DenseNet-121 to detect the fine-grained spatial heterogeneity within the tumor tissue. Small patches (224 × 224 pixels) were extracted from the whole-slide images (~10,000 × 10,000

pixels) centered on the spatial transcriptomics spots, and each patch went through the pretrained DenseNet-121 network for training and predicting the preselected 250 genes with the highest mean expression.

Performance assessment

To assess the performance of models in Application 1, we adopted metrics including accuracy and macro-F1, which are defined below

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n} \text{Precision}_i$$

$$\text{Recall}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n} \text{Recall}_i$$

$$\text{F1}_{\text{macro}} = \frac{2 * \text{Precision}_{\text{macro}} * \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, y_i is the predicted value, $\widetilde{y}_i$ is the real value, and n is the total number of samples.

To evaluate PPML-Omics on Application 2, we applied clustering ARI score, NMI, CA, and JI to evaluate our method's utility comprehensively. The evaluation metrics between the predicted cluster B with the ground truth cluster A in ARI, NMI, CA, and JI are defined below

$$\text{ARI} = \frac{\sum_{ij}\binom{n_i j}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

$$\text{NMI}(A, B) = \frac{2 * I(A; B)}{H(Y) + H(C)}$$

$$\text{CA}(A, B) = \max_{\text{perm} \in P} \frac{1}{n}\sum_{i=1}^{n-1} \left[\text{perm}(B_i) = A_i\right]$$

$$\text{JI}(A, B) = \frac{A \bigcap B}{A \cup B}$$

where A is the class label, B is the cluster label, $H(\cdot)$ is the entropy, and $I(A; B)$ is the mutual information between A and B. In the definition of CA, P stands for the set of all permutations in $[1 : K]$ where K is the number of clusters, and n is the sample size.

To measure the similarity between the predicted gene expression and the ground truth in Application 3, we used the MSE as the evaluation metric

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Implementation of federated learning framework

As shown in Algorithm 1, the FL framework aimed at ensuring data privacy and security along with the improvement of the AI model based on joint data sources from multiple clients around the world. We first initiated multiple data sources as our clients. Since the essence of FL is the union of samples, each client first needs to download the model from the server and initiate their client model with the server weight. Then, each participant can use local data to train the model, calculate their gradient, and upload it to the server. The server needs to aggregate the gradient of each client to update the server model parameters. In this process, each client is treated with the same and complete model, and there is no communication and no dependence among clients. Therefore, each client can also make independent predictions during the prediction.

Algorithm 1 Federated Learning

Input: K is the number of clients with data source $D_1, D_2, \ldots, D_K$; T is the number of epochs; η is learning rate; C is the parameter for l2-norm clipping;

Procedure Server Execution:

Initialize: w_0
For $t = 1, \ldots, T$ **do**
  For each client $k : 1, \ldots, K$ in parallel **do**
    $\Delta w_{t+1}^k \leftarrow \text{ClientUpdate}(w_t, D_k)$
    $\Delta \widetilde{w}_{t+1}^k \leftarrow \Delta w_{t+1}^k / \max\left(1, \frac{\|\Delta w_{t+1}^k\|_2}{C}\right)$
  $w_{t+1} \leftarrow w_t + \eta\left(\frac{1}{K}\sum_{k=1}^{K} \Delta \widetilde{w}_{t+1}\right)$

function ClientUpdate(w_t, D_k)
  $w \leftarrow w_t$
  For each patient $p \in D_k$ **do**
    $w \leftarrow w - \eta \nabla \ell(w, p)$
  $\Delta w_{t+1} = (w - w_t)/\eta$
  return Δw_{t+1}

Implementation of differential private model training

To add noise to the gradient, we used the calibrated analytic Gaussian mechanism to calculate the value of σ in the Gaussian distribution based on the values of ϵ and δ. In each epoch, after computing the gradient update for each client, we clipped the gradient to ensure a finite upper and lower bound and then added noise to each value in the gradient that fitted the previously defined Gaussian distribution as shown in Algorithm 2. A common paradigm for approximating a deterministic real-valued function $f : D \to R$ with Gaussian differentially private mechanism is via additive noise calibrated to f's sensitivity S_f, which is defined as the maximum of the l_2 norm $\|f(D) - f(D')\|_2$ where D and D' are neighboring inputs. Proving the DP guarantee in the SGD algorithm requires bounding the influence of each sample on the gradient. Since there was no prior bound on the size of the gradients, we clipped each gradient in the l_2 norm by C.

Implementation of DR protocol

In a practical scenario, before the end of each epoch, each client sends the gradient information to the server, while the server knows where each gradient information comes from, providing the possibility of privacy leakage. With DR protocol, we can still ensure that the gradients received by the server are randomly shuffled. As shown in Algorithm 2, in each epoch of training, after the DR mechanism, all clients may not hold their own original gradients, but rather gradients from a randomly paired client. Then, all clients upload their gradients to the server.

σ Selection with analytic Gaussian DP mechanism

To achieve end-to-end (ϵ_T, δ_T)-DP for T iterations in total with PPML-Omics, we need to achieve (ϵ_c, δ_c)-DP for the server in each epoch where $\epsilon_c = \frac{\epsilon_T}{2\sqrt{2Tln(2/\delta_T)}}$ and $\delta_c = \frac{\delta_T}{2T}$, and further achieve ($\epsilon_l$, δ_l)-DP for each client at each epoch, where $\epsilon_l = \frac{\epsilon_c\sqrt{K}}{\sqrt{ln(2/\delta_c)}}$ and $\delta_l = \frac{\delta_c}{K}$. With the values of ϵ_l, δ_l, we could calculate the standard deviation σ of Gaussian noise needed in each epoch to each gradient transmitted from each client by using the analytic Gaussian mechanism of (*78*), where σ = CalibrateAnalyticGaussianMechanism(ϵ_l, δ_l). For FL-DP method, we could calculate $\epsilon_l = \frac{\epsilon_T}{2\sqrt{2Tln(1/\delta_T)}}$ and $\delta_l = \frac{\delta_T}{2T}$.

Let $f: X^n \to \mathbb{R}^k$ be a function with global l_2-norm sensitivity Δ. Suppose $\epsilon > 0$ and $0 < \delta < \frac{1}{2} - e^{-\frac{3\epsilon}{\sqrt{4\pi\epsilon}}}$. If the Gaussian mechanism $A(D) = f(D) + Z$ with $Z \sim N(0,\sigma^2\mathbb{I}_k)$ is (ϵ, δ)-DP, then $\sigma \geq \frac{\Delta}{\sqrt{2\epsilon}}$. When $\sigma = \frac{\Delta}{\sqrt{2\epsilon}}$ in analytic Gaussian DP mechanism, from theorem 1 in the Supplementary Materials, the mechanism will be (ϵ, δ)-DP with $\delta = \Phi(0) - e^{\epsilon}\Phi(-\sqrt{2\epsilon}) > \frac{1}{2} - e^{-\frac{3\epsilon}{\sqrt{4\pi\epsilon}}}$. Thus, it is impossible to achieve (ϵ, δ)-DP with $\delta = \Phi(0) - e^{\epsilon}\Phi(-\sqrt{2\epsilon})$ without increasing the variance of perturbation. We could use the above theorem to get the minimal noise to add for the given ϵ_l for each client at each epoch.

Algorithm 2 PPML-Omics

Input: K is the number of clients with data source $D_1, D_2, \ldots, D_K$; T is the number of epochs; d_k is the number of samples of client k; d is the total number of samples of all clients; ϕ_k is the weight of gradient for client k; C is the parameter for l_2-norm clipping; η is learning rate; $DR(k)$ is the id of the owner of $\Delta\widetilde{w}_{t+1}^{DR(k)}$ after executing DecentralizedRandomization on the original owner k with $\Delta\widetilde{w}_{t+1}^{k}$; In order to achieve end-to-end (ϵ_T,δ_T)-DP for T iterations in total with PPML-Omics, we need to achieve (ϵ_c, δ_c)-DP for the server in each epoch where $\epsilon_c = \frac{\epsilon_T}{2\sqrt{2Tln(2/\delta_T)}}$ and $\delta_c = \frac{\delta_T}{2T}$, and further achieve ($\epsilon_l$,$\delta_l$)-DP for each client at each epoch, where $\epsilon_l = \frac{\epsilon_c\sqrt{K}}{\sqrt{ln(2/\delta_c)}}$ and $\delta_l = \frac{\delta_c}{K}$. With the values of ϵ_l and δ_l, we could calculate the standard deviation σ of Gaussian noise needed in each epoch to each gradient transmitted from each client by using the analytic Gaussian mechanism of (*77*), where $\sigma = CalibrateAnalyticGaussianMechanism(\epsilon_l, \delta_l)$.

Procedure Server Execution:

Initialize: w_0
For $t = 1, \ldots, T$ **do**
  For each client $k: 1, \ldots, K$ in parallel **do**
    $\Delta w_{t+1}^k \leftarrow$ ClientUpdate(w_t, D_k)
    $\Delta\widetilde{w}_{t+1}^k \leftarrow \Delta w_{t+1}^k / \max\left(1, \frac{\|\Delta w_{t+1}^k\|_2}{C}\right) + \mathcal{N}(0, \sigma^2\boldsymbol{I})$
  DecentralizedRandomization(K)
  Collect gradients from all clients: $\left(\Delta\widetilde{w}_{t+1}^{DR(k)}\right)_{k\in[K]}$
  $w_{t+1} \leftarrow w_t + \eta\left(\sum_{k=1}^{K}\Delta\widetilde{w}_{t+1}^{DR(k)}\right)$

function ClientUpdate(w_t, D_k)
  $w \leftarrow w_t$
  For each patient $p \in D_k$ **do**
    $w \leftarrow w - \eta\nabla\ell(w, p)$
  $\phi_k = d_k/d$
  $\Delta w_{t+1} = \phi_k(w - w_t)/\eta$
  return Δw_{t+1}

function DecentralizedRandomization(K)
  For $i: 1, \ldots, K-1$ **do**
    $j \leftarrow$ random integer such that $i \leq j \leq K$
    client i sends $\Delta\widetilde{w}_{t+1}^i$ to client j
    client j sends $\Delta\widetilde{w}_{t+1}^j$ to client i
    client i and j use the received $\Delta\widetilde{w}_{t+1}^j$ and $\Delta\widetilde{w}_{t+1}^i$ to overwrite the local $\Delta\widetilde{w}_{t+1}^i$ and local $\Delta\widetilde{w}_{t+1}^j$

Hyperparameter selection in DP mechanism

Regarding the parameters in the DP mechanism, we selected the proper value **C** for l_2-norm clipping based on statistically observing the distribution of all elements in gradients during the training phase for the centrally trained model, thus estimating the sensitivity of the training procedure of PPML-Omics. The selection of the privacy budget ϵ was a tricky task, as commonly adopted, such as in the handbook "Disclosure Avoidance for the 2020 Census: An Introduction" (*59*) and in (*86*), we tested ϵ from 0.1, 0.5, 1, 5, 10, 20, 30, 40, 50, and measured the utility against different end-to-end ϵ_T for all applications as a reference the potential users of PPML-Omics. To choose the proper value of δ, we performed the grid search from 0 to $1/K$, where K is the number of clients.

Hyperparameter and model selection in deep learning

To ensure that all models had a fair chance of learning a useful representation in all tasks, we trained multiple instances of each model (FCN for Application 1, Auto-encoder for Application 2, and ST-Net for Application 3) using a grid search of hyperparameter settings, including learning rate, epochs, number of hidden layers, number of filters, filter size, batch size, momentum, initial weight, weight decay, and keep probability. Then, we selected model instances based on their training performance.

Model inversion attack for cancer classification models

ML model was abused to learn sensitive genomic information about individuals, which was shown in a case study of linear classifiers in personalized medicine by Fredrikson *et al.* (*87*). Thus, we implemented the MIA proposed by Fredrikson *et al.* (*88*) to extract sensitive information from the trained models with our PPML-Omics method under different privacy budgets to address the privacy concerns. Regarding our tasks, we retreated the MIA as an optimization problem: to find the input that maximizes the returned class confidence score (*88*), which was achieved by using gradient descent along with modifications specific to this domain as shown in Algorithm 3 (section S2.3).

We reconstructed the original gene expression with the MIA for each cancer for all pretrained models. For each cancer, we applied MIA to reconstruct the gene expression vector for the targeting cancer and selected genes with the highest reconstructed expression levels (≥0.8) as the most significant reconstructed gene features specific to each cancer. Then, we split all samples of 37 cancers into two groups (one only included the samples of the targeting cancer type and the other one included all remaining samples from other cancer types) and plotted the distribution of the z-score normalized expression of the previously selected genes of the real expression data in both groups (solid lines for samples of the target cancer type and dashed lines for other cancer types).

Algorithm 3 Model Inversion Attack (MIA)

Input: Pre-trained model M, target label $\boldsymbol{t}$, aim input to optimize $\boldsymbol{i}$, number of epochs α, learning rate η, early stopping parameter β and cost threshold γ.

Procedure:

Initialize $\boldsymbol{i}$ with 0
For epoch in {0 ... α} **do**
  Get prediction $\boldsymbol{o} = M(\boldsymbol{i})$
  Calculate cost $c = LossFunction(\boldsymbol{o}, \boldsymbol{t})$
  Calculate gradient Δ on c
  Update $\boldsymbol{i} = \boldsymbol{i} - \eta * \Delta$
  if c keeps increasing for β epochs **then** Halt
  if $c \leq \gamma$ **then** Halt

iDLG reconstruction attack

Sharing gradients during the training of ML networks could potentially leak private data. Zhao *et al.* (*77*) presented an approach named iDLG as shown in Algorithm 4, which showed the possibility to obtain private training data from the publicly shared gradients. In integrating tumor morphology and gene expression with spatial transcriptomics, we synthesized the dummy data and corresponding labels with the supervision

of shared gradients and optimized the dummy data with iDLG to obtain one similar to the private training data.

Algorithm 4 iDLG

Input: Differentiable model $\boldsymbol{M}$, model parameters $\boldsymbol{W}$, private training data and labels $(\boldsymbol{x}, \boldsymbol{c})$, gradients produced by the private data $\nabla \boldsymbol{W}$,dummy data and labels $(\boldsymbol{x}', \boldsymbol{c}')$, number of iterations N, learning rate η and loss function l.

Procedure:

Extract the target ground-truth label to initialize the dummy label $\boldsymbol{c}'$
Initialize the dummy data $\boldsymbol{x}' \leftarrow \mathcal{N}(0,1)$
For i ← 1 to N **do**
  Calculate the dummy gradients: $\nabla \boldsymbol{W} \leftarrow \partial l(\boldsymbol{M}(\boldsymbol{x}', \boldsymbol{W}), c')/\partial \boldsymbol{W}$
  Calculate the loss: $L_G = ||\nabla \boldsymbol{W}' - \nabla \boldsymbol{W}||_F^2$
  Update the dummy data: $\boldsymbol{x}' \leftarrow \boldsymbol{x}' - \eta \nabla_{x'} L_G$

Quantification of privacy leakage

For Application 1, to investigate the significance of the results of gene reconstruction attacks, we used the Kullback-Leibler divergence, JS divergence, and KS test as shown below to characterize the degree of privacy leakage by calculating the divergence between two distributions of the expression level of the reconstructed genes in the target cancer type and other cancer types. Larger JS divergence values and smaller P values of the KS test indicated more significant differences between the expression levels of the reconstructed signature genes in the target cancer type and other cancer types, suggesting a severe privacy leakage

$$\mathrm{KL}(P \,||\, Q) = \sum p(x) \log \frac{P(x)}{Q(x)}$$

$$JS(P \,||\, Q) = \frac{1}{2}\mathrm{KL}\left[P(x) \,||\, \frac{P(x)+Q(x)}{2}\right] + \frac{1}{2}\mathrm{KL}\left[Q(x) \,||\, \frac{P(x)+Q(x)}{2}\right]$$

where $P(x)$ and $Q(x)$ are two probability distributions.

For Application 2, given a method, if we cannot observe some targeted small clusters (sub-types) in the final clustering and the proportion results, then we could conclude that the method protects the patient's privacy. In addition, if the clustering results for major cell types are correct, we could conclude that the method preserves a reasonable degree of usability while protecting privacy.

PPML-Omics avoids reconstruction attacks on medical images. Thus, we evaluated the privacy-preserving capability by visually comparing the reconstructed image and the raw image after the iDLG attack on different methods in Application 3.

Multi-party homomorphic encryption

To compare PPML-Omics (DP-based solution) with the solution in the cryptographic track, we implemented an FL method with MHE based on the CKKS cryptographic scheme (*61*) that provides approximate arithmetic over vectors of complex numbers in the training phase of the FL method with TenSEAL (*89*), which is a library for doing HE operations on tensors, built on top of Microsoft SEAL.

Supplementary Materials

This PDF file includes:

Sections S1 to S5
Figs. S1 to S11
Tables S1 to S12
References

REFERENCES AND NOTES

1. Y. Joly, S. O. M. Dyke, B. M. Knoppers, T. Pastinen, Are data sharing and privacy protection mutually exclusive? *Cell* **167**, 1150–1154 (2016).
2. K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp Oncol* **19**, A68–A77 (2015).
3. G. England, The 100,000 genomes project protocol v3 genomics England. *Genomics Engl. Protoc.*, (2017).
4. H. A. Lewin, G. E. Robinson, W. J. Kress, W. J. Baker, J. Coddington, K. A. Crandall, R. Durbin, S. V. Edwards, F. Forest, M. T. P. Gilbert, Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* **115**, 4325–4333 (2018).
5. M. Alser, H. Hassan, H. Xin, O. Ergin, O. Mutlu, C. Alkan, GateKeeper: A new hardware architecture for accelerating pre-alignment in DNA short read mapping. *Bioinformatics* **33**, 3355–3363 (2017).
6. Y. Joly, I. N. Feze, L. Song, B. M. Knoppers, Comparative approaches to genetic discrimination: Chasing shadows? *Trends Genet.* **33**, 299–302 (2017).
7. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
8. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
9. V. Marx, Method of the year: Spatially resolved transcriptomics. *Nat. Methods* **18**, 9–14 (2021).
10. J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, A. Telenti, A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
11. G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell* **2**, 305–311 (2020).
12. A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
13. M. Al-Rubaie, J. M. Chang, Privacy-preserving machine learning: Threats and solutions. *IEEE Secur. Priv.* **17**, 49–58 (2019).
14. M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **10**, 12598 (2020).
15. C. G. Schwarz, W. K. Kremers, T. M. Therneau, R. R. Sharp, J. L. Gunter, P. Vemuri, A. Arani, A. J. Spychalla, K. Kantarci, D. S. Knopman, Identification of anonymous MRI research participants with face-recognition software. *NEJM* **381**, 1684–1686 (2019).
16. A. Harmanci, M. Gerstein, Quantification of private information leakage from phenotype-genotype data: Linking attacks. *Nat. Methods* **13**, 251–256 (2016).
17. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in *Artificial Intelligence and Statistics*. (PMLR, 2017), pp. 1273–1282.
18. B. Hitaj, G. Ateniese, F. Perez-Cruz, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), Dallas, TX, 30 October to 3 November, pp. 603–618.
19. L. Melis, C. Song, E. De Cristofaro, V. Shmatikov, *2019 IEEE Symposium on Security and Privacy* (*SP*) San Francisco, CA, 20 to 22 May 2019, (IEEE, 2019), pp. 691–706.
20. M. Nasr, R. Shokri, A. Houmansadr, *2019 IEEE Symposium on Security and Privacy* (*SP*) San Francisco, CA, 20 to 22 May 2019, (IEEE, 2019), pp. 739–753.
21. L. Zhu, Z. Liu, S. Han, Deep leakage from gradients. *Adv. Neural Inf. Process.* **32**, (2019).
22. V. Tolpegin, S. Truex, M. E. Gursoy, L. Liu, *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020*, Guildford, UK, 14 to 18 September 2020, Proceedings, Part I 25. (Springer, 2020), pp. 480–501.
23. M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, Y. Wang, Membership inference attack against differentially private deep learning model. *Trans. Data Priv.* **11**, 61–79 (2018).
24. A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, M. Backes, Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv:1806.01246, (2018).
25. H. Hu, Z. Salcic, L. Sun, G. Dobbie, X. Zhang, *2021 IEEE International Conference on Data Mining* (*ICDM*) Auckland, New Zealand, 7 to 10 December 2021, (IEEE, 2021), pp. 1102–1107.
26. J. Geiping, H. Bauermeister, H. Dröge, M. Moeller, Inverting gradients-how easy is it to break privacy in federated learning? *Adv. Neural Inf. Process.* **33**, 16937–16947 (2020).
27. H. J. La, M. K. Kim, S. D. Kim, 2015 *IEEE International Conference on Services Computing* New York City, NY, 27 June to July 2 2015, (IEEE, 2015), pp. 249–255.
28. K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, K. Seth, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), Dallas, TX, 30 October to 3 November 2017, pp. 1175–1191.
29. S. Wagh, D. Gupta, N. Chandran, SecureNN: 3-Party secure computation for neural network Training. *PoPETs* **2019**, 26–49 (2019).
30. S. Sav, A. Pyrgelis, J. R. Troncoso-Pastoriza, D. Froelicher, J.-P. Bossuat, J. S. Sousa, J.-P. Hubaux, POSEIDON: Privacy-preserving federated neural network learning. arXiv:2009.00349 (2020).

31. D. Froelicher, J. R. Troncoso-Pastoriza, J. L. Raisaro, M. A. Cuendet, J. S. Sousa, H. Cho, B. Berger, J. Fellay, J.-P. Hubaux, Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat. Commun.* **12**, 5910 (2021).
32. S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N. A. Aziz, Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
33. M. Ali, H. Karimipour, M. Tariq, Integration of blockchain and federated learning for Internet of Things: Recent advances and future challenges. *Comput. Secur.* **108**, 102355 (2021).
34. W. Huang, M. Zhuo, T. Zhu, S. Zhou, Y. Liao, Differential privacy: Review of improving utility through cryptography-based technologies. *Concurr. Comput.* **35**, e7565 (2023).
35. C. Dwork, Differential privacy: A survey of results, *International Conference on Theory and Applications of Models of Computation* (Springer, 2008), pp. 1–19.
36. K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, H. V. Poor, Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Trans. Inf. Forensics Secur.* **15**, 3454–3469 (2020).
37. N. Rodríguez-Barroso, G. Stipcich, D. Jiménez-López, J. A. Ruiz-Millán, E. Martínez-Cámara, G. González-Seco, M. V. Luzón, M. A. Veganzones, F. Herrera, Federated learning and differential privacy: Software tools analysis, the Sherpa. ai FL framework and methodological guidelines for preserving data privacy. *Inf. Fusion* **64**, 270–292 (2020).
38. R. Liu, Y. Cao, H. Chen, R. Guo, M. Yoshikawa, *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), Vancouver Convention Centre, Vancouver, Canada, 2 to 9 February 2021, vol. 35, pp. 8688–8696.
39. A. Girgis, D. Data, S. Diggavi, P. Kairouz, A. T. Suresh, *International Conference on Artificial Intelligence and Statistics*. Virtual Conference, 13 to 15 April 2021, (PMLR, 2021), pp. 2521–2529.
40. B. Ghazi, R. Kumar, P. Manurangsi, R. Pagh, A. Sinha, *International Conference on Machine Learning*. Virtual Conference, 18 to 24 July 2021, (PMLR, 2021), pp. 3692–3701.
41. S. D. Constable, Y. Tang, S. Wang, X. Jiang, S. Chapin, Privacy-preserving GWAS analysis on federated genomic datasets, in *BMC Medical Informatics and Decision Making*. (BioMed Central, 2015), vol. 15, pp. 1–9.
42. H. Cho, D. J. Wu, B. Berger, Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
43. G. Gürsoy, E. Chielle, C. M. Brannon, M. Maniatakos, M. Gerstein, Privacy-preserving genotype imputation with fully homomorphic encryption. *Cell Systems* **13**, 173–182 (2022).
44. Z. He, Y. Li, J. Li, K. Li, Q. Cai, Y. Liang, Achieving differential privacy of genomic data releasing via belief propagation. *TST* **23**, 389–395 (2018).
45. E. Yilmaz, T. Ji, E. Ayday, P. Li, Genomic data sharing under dependent local differential privacy. *CODASPY* **2022**, 77–88 (2022).
46. N. Almadhoun, E. Ayday, O. Ulusoy, Differential privacy under dependent tuples-the case of genomic privacy. *Bioinformatics* **36**, 1696–1703 (2020).
47. G. Gursoy, C. M. Brannon, F. C. P. Navarro, M. Gerstein, FANCY: Fast estimation of privacy risk in functional genomics data. *Bioinformatics* **36**, 5145–5150 (2021).
48. G. Gursoy, T. Li, S. Liu, E. Ni, C. M. Brannon, M. B. Gerstein, Functional genomics data: Privacy risk assessment and technological mitigation. *Nat. Rev. Genet.* **23**, 245–258 (2022).
49. O. Zolotareva, R. Nasirigerdeh, J. Matschinske, R. Torkzadehmahani, M. Bakhtiari, T. Frisch, J. Spath, D. B. Blumenthal, A. Abbasinejad, P. Tieri, G. Kaissis, D. Ruckert, N. K. Wenke, M. List, J. Baumbach, Flimma: A federated and privacy-aware tool for differential gene expression analysis. *Genome Biol.* **22**, 338 (2021).
50. M. M. Islam, N. Mohammed, Y. Wang, P. Hu, Differential private deep learning models for analyzing breast cancer omics data. *Front. Oncol.* **12**, 879607 (2022).
51. W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso, A. Feng, Privacy-preserving federated brain tumour segmentation, in *Machine Learning in Medical Imaging*, H.-I. Suk, M. Liu, P. Yan, C. Lian, Eds. (Springer International Publishing, 2019), pp. 133–141.
52. G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, R. Braren, End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nat. Mach. Intell.* **3**, 473–484 (2021).
53. M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, H. R. Tizhoosh, Federated learning and differential privacy for medical image analysis. *Sci. Rep.* **12**, 1953 (2022).
54. E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
55. D. J. Burgess, Spatial transcriptomics coming of age. *Nat. Rev. Genet.* **20**, 317 (2019).
56. B. He, L. Bergenstrahle, L. Stenbeck, A. Abid, A. Andersson, A. Borg, J. Maaskola, J. Lundeberg, J. Zou, Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
57. R. A. Fisher, F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research* (Hafner Publishing Company, 1953).
58. T. Li, N. Li, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009.
59. United States Census Bureau. Disclosure avoidance for the 2020 census: An introduction. (2020).
60. V. Feldman, I. Mironov, K. Talwar, A. Thakurta, *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, Paris, France, 7 to 9 October 2018.
61. J. H. Cheon, A. Kim, M. Kim, Y. Song, Homomorphic encryption for arithmetic of approximate numbers, in *Advances in Cryptology–ASIACRYPT 2017*, T. Takagi, T. Peyrin, Eds. (Springer International Publishing, Cham, 2017), pp. 409–437.
62. S. Cristea, K. Polyak, Dissecting the mammary gland one cell at a time. *Nat. Commun.* **9**, 2473 (2018).
63. L. González-Silva, L. Quevedo, I. Varela, Tumor functional heterogeneity unraveled by scRNA-seq technologies. *Trends in cancer* **6**, 13–19 (2020).
64. Q. Zhang, J. Wang, P. Wang, T. Tang, P. Li, Y. Pei, X. Zhang, W. Zhang, Q. Gu, Q. Ji, Establishment and optimization of scrna-seq assay to find the mechanism of immune therapy against tumors. *Cell* **8**, 9 (2021).
65. S. Ding, X. Chen, K. Shen, Single-cell RNA sequencing in breast cancer: Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer Commun.* **40**, 329–344 (2020).
66. D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, T. Nguyen, Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.* **12**, 1029 (2021).
67. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
68. V. Y. Kiselev, K. Kirschner, M. T. Schaub, T. Andrews, A. Yiu, T. Chandra, K. N. Natarajan, W. Reik, M. Barahona, A. R. Green, SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
69. P. Lin, M. Troup, J. W. Ho, CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 1–11 (2017).
70. M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, Y. Xu, SINCERA: A pipeline for single-cell RNA-Seq profiling analysis. *PLoS Comput. Biol.* **11**, e1004575 (2015).
71. L. Zhang, Z. Li, K. M. Skrzypczynska, Q. Fang, W. Zhang, S. A. O'Brien, Y. He, L. Wang, Q. Zhang, A. Kim, Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell* **181**, 442-459. e429 (2020).
72. Y. Zeng, Z. Wei, W. Yu, R. Yin, Y. Yuan, B. Li, Z. Tang, Y. Lu, Y. Yang, Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.* **23**, bbac297 (2022).
73. C. Song, T. Ristenpart, V. Shmatikov, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (2017), Dallas, TX, 30 October to 3 November 2017, pp. 587–601.
74. A. Raj, Y. Bresler, B. Li, *International Conference on Machine Learning*. Virtual conference, 12 to 18 July 2020, (PMLR, 2020), pp. 7932–7942.
75. R. Cappelli, D. Maio, A. Lumini, D. Maltoni, Fingerprint image reconstruction from standard templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1489–1503 (2007).
76. Z. Li, M. Hubchak, Y. Zhu, *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. Victoria, BC, Canada, 9 to 12 August 2021, (IEEE, 2021), pp. 447–448.
77. B. Zhao, K. R. Mopuri, H. Bilen, idlg: Improved deep leakage from gradients. Stockholm, Sweden, 10 to 15 July 2018, [Preprint] arXiv:2001.02610 (2020).
78. B. Balle, Y.-X. Wang, *International Conference on Machine Learning*. San Diego, CA, 6 to 9 January 2019, (PMLR, 2018), pp. 394–403.
79. Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, A. Thakurta, *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (SIAM, 2019), San Diego, CA, 6 to 9 January 2019, pp. 2468–2479.
80. X. Li, M. Jiang, X. Zhang, M. Kamp, Q. Dou, Fedbn: Federated learning on non-iid features via local batch normalization. arXiv:2102.07623 (2021).
81. M. Andreux, J. O. du Terrail, C. Beguier, E. W. Tramel, *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 2* (Springer, 2020), pp. 129–139.
82. L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
83. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
84. S. Hrvatin, D. R. Hochbaum, M. A. Nagy, M. Cicconet, K. Robertson, L. Cheadle, R. Zilionis, A. Ratner, R. Borges-Monroy, A. M. Klein, Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
85. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process.* **32**, (2019).
86. B. Jayaraman, D. Evans, *28th USENIX Security Symposium* (*USENIX Security 19*) (2019), Santa Clara, CA, 14 to 16 August 2019, pp. 1895–1912.

87. M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, *23rd USENIX Security Symposium* (*USENIX Security 14*) (2014), San Diego, CA, 20 to 22 August 2014, pp. 17–32.
88. M. Fredrikson, S. Jha, T. Ristenpart, *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015), Denver, CO, 12 to 16 October 2015, pp. 1322–1333.
89. A. Benaissa, B. Retiat, B. Cebere, A. E. Belfedhal, TenSEAL: A library for encrypted tensor operations using homomorphic encryption. arXiv:2104.03152 (2021).
90. L. Sweeney, Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* **25**, 98–110 (1997).
91. A. Narayanan, V. Shmatikov, *2008 IEEE Symposium on Security and Privacy* (*sp 2008*). Oakland, CA, 18 to 21 May 2008, (IEEE, 2008), pp. 111–125.
92. C. Dwork, F. McSherry, K. Nissim, A. Smith, *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006*, New York, NY, 4 to 7 March 2006 (Springer, 2006), pp. 265–284.
93. C. Dwork, A. Roth, The algorithmic foundations of differential privacy. *Theor. Comput. Sci.* **9**, 211–407 (2014).
94. A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. R. O'Brien, T. Steinke, S. Vadhan, Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.* **21**, 209 (2018).
95. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, FL, 20 to 25 June 2009, (IEEE, 2009), pp. 248–255.
96. A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, G. Kaissis, Medical imaging deep learning with differential privacy. *Sci. Rep.* **11**, 13524 (2021).
97. Y. Erlich, A. Narayanan, Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014).
98. A. Johnson, V. Shmatikov, *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2013), Chicago, IL, 11 to 14 August 2013, pp. 1079–1087.
99. C. Uhlerop, A. Slavković, S. E. Fienberg, Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.* **5**, 137–166 (2013).
100. F. Yu, S. E. Fienberg, A. B. Slavković, C. Uhler, Scalable privacy-preserving data sharing methodology for genome-wide association studies. *J. Biomed. Inform.* **50**, 133–141 (2014).
101. D. Grishin, J. L. Raisaro, J. R. Troncoso-Pastoriza, K. Obbad, K. Quinn, M. Misbach, J. Gollhardt, J. Sa, J. Fellay, G. M. Church, Citizen-centered, auditable and privacy-preserving population genomics. *Nat. Comput. Sci.* **1**, 192–198 (2021).
102. J. L. Raisaro, G. Choi, S. Pradervand, R. Colsenet, N. Jacquemont, N. Rosat, V. Mooser, J.-P. Hubaux, Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 1413–1426 (2018).
103. C.-A. Azencott, Machine learning and genomics: Precision medicine versus patient privacy. *Philos. Trans. Royal Soc. A* **376**, 20170350 (2018).
104. B. Balle, J. Bell, A. Gascón, K. Nissim, *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference,* Santa Barbara, CA, 18 to 22 August 2019, Proceedings, Part II 39 (Springer, 2019), pp. 638–667.
105. P. Kairouz, S. Oh, P. Viswanath, The composition theorem for differential privacy, in *International Conference on Machine Learning*. (PMLR, 2015), pp. 1376–1385.
106. B. Li, C. N. Dewey, RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform* **12**, 1–16 (2011).
107. A.-C. Villani, R. Satija, G. Reynolds, S. Sarkizova, K. Shekhar, J. Fletcher, M. Griesbeck, A. Butler, S. Zheng, S. Lazo, Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).

Acknowledgments

Funding: J.Z., S.C., H.L., B.Z., L.Z., Z.L., N.C., W.H., C.X., and X.G. were supported in part by grants from Office of Research Administration (ORA) at King Abdullah University of Science and Technology (KAUST) under award numbers URF/1/4352-01-01, FCC/1/1976-44-01, FCC/1/1976-45-01, REI/1/5234-01-01, REI/1/5414-01-01, REI/1/5289-01-01, REI/1/5404-01-01. Y.W., Y.H., Z.X., and D.W. were supported in part by the baseline funding BAS/1/1689-01-01 and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). **Author contributions:** Conceptualization: J.Z., S.C., and X.G. Methodology: J.Z., S.C., Y.W., X.G., and D.W. Investigation: J.Z., S.C., and Y.W. Visualization: J.Z., S.C., H.L., B.Z., Z.L., and N.C. Supervision: X.G. and D.W. Writing—original draft: J.Z., S.C., and Y.W. Writing—review and editing: J.Z., S.C., Y.W., H.L., B.Z., L.Z., Y.H., Z.X., Z.L., N.C., W.H., D.W., C.X., and X.G. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data for the cancer classification with bulk RNA-seq are based on the one generated by the TCGA Research Network: www.cancer.gov/tcga. The datasets for scRNA-seq clustering analysis are available at GEO accession ID or Sequence Read Archive: Yan (GSE36552), Pollen (SRP041736), and Hrvatin (GSE102827). The postprocessed datasets for scRNA-seq clustering analysis are also available at https://zenodo.org/records/10198453. scRNA-seq data of patients are available at GEO accession ID: GSE146771. For the spatial transcriptomics dataset, there are 30,612 spots, including 23 patients with breast cancer with four subtypes: luminal A, luminal B, triple-negative, and HER2-positive. There were three high-resolution images of H&E staining tissue and their corresponding gene expression data for each patient. All images and processed data are available at https://data.mendeley.com/datasets/29ntw7sh4r/5. All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Code for this paper is available at https://github.com/JoshuaChou2018/PPML-Omics and https://doi.org/10.5281/zenodo.8247563. For more in-depth information on hyperparameters, please consult the code repository.

Submitted 18 March 2023
Accepted 29 December 2023
Published 31 January 2024
10.1126/sciadv.adh8601