Efficient Federated Kinship Relationship Identification

Xinyue Wang[1], Leonard Dervishi[2], Wentao Li[3], Xiaoqian Jiang[3], Erman Ayday[2], Jaideep Vaidya[1]
[1] Rutgers University, Newark, NJ; [2]Case Western Reserve University, Cleveland, OH; [3]UTHealth, Houston, TX

Abstract

Kinship relationship estimation plays a significant role in today's genome studies. Since genetic data are mostly stored and protected in different silos, retrieving the desirable kinship relationships across federated data warehouses is a non-trivial problem. The ability to identify and connect related individuals is important for both research and clinical applications. In this work, we propose a new privacy-preserving kinship relationship estimation framework: Incremental Update Kinship Identification (INK). The proposed framework includes three key components that allow us to control the balance between privacy and accuracy (of kinship estimation): an incremental process coupled with the use of auxiliary information and informative scores. Our empirical evaluation shows that INK can achieve higher kinship identification correctness while exposing fewer genetic markers.

Introduction

In the age of big data, collaborative studies have increasingly become more important. Especially in genomic research, all the participating parties in federated setting benefit from such collaborations since they obtain more significant statistics and a more accurate outcome [1, 2]. On the other hand, genomic datasets typically contain sensitive information of the participants such as phenotype, family membership, and disease information, which, if inferred by an adversary, may be used in a harmful way (e.g., higher health insurance rates or unemployment due to known preconditions) [3, 4]. Some of the existing solutions that mitigate such privacy issues include: (a) differential privacy based solutions, which generally rely on addition of noise to the raw data [5, 6]; (b) meta-analysis, in which the collaborators only share the aggregate statistics of individual studies with each other [7, 8]; and (c) cryptographic solutions, which utilize cryptographic algorithms to encrypt the data, but also enable the researchers to work on encrypted data [9, 10].

Often the datasets that are being used in genomic research may introduce bias to the research results. This is because the samples may be related to each other, or the majority of the samples may belong to only one or two sub-populations. This could influence the results of the research, and for any study it is important to be aware of this potential bias. Typically since any study would primarily be focused on a particular target group, one of the most important steps in collaborative studies is to select the set of attributes and samples that will be used in the study. This process helps to ensure that the results of the study are fair and robust.

Our focus in this work is the identification of the related samples in the federated dataset of the collaborators. This is typically required in several scenarios: (a) the collaborators identify and filter out the related samples as part of the quality control procedure which is widely used in genome-wide association studies (GWAS); or (b) collaborators identify closely related samples (family members) and use only those samples to perform GWAS. Note that the related samples may belong to datasets owned by different parties and that is the main challenge we address in this paper.

Previous works in this area have mainly focused on the privacy-preserving similar patient search, which means identifying similar patients in the datasets to a known target. For example, Jha et al. [11], proposed a privacy-preserving cryptographic technique for computing the Smith-Waterman similarity score and edit distance between two sequences. Recently, Zhu et al. [12] proposed a secure patient search mechanism based on a gBK-tree, edit distance approximation algorithm, and symmetric-key encryption. As opposed to the aforementioned approaches, our work focuses on a different problem of finding similar records across different genomic datasets without a known target. The scheme proposed in [13] is the first to provide a privacy-preserving solution to identify related individuals across different genomic datasets. In the proposed framework, the collaborators synchronize to decide on a set of genetic markers (single nucleotide polymorphisms - SNPs) and a common seed, shuffle their data according to the seed, and send their shuffled dataset to the server which performs the necessary computations to compute the kinship coefficients between all the federated data participants. With more than 250 SNPs, their method correctly identifies 95% of kinship rela-

tionships. However, the authors only consider up to second degree relationships. Moreover, due to the usage of a local differential privacy (ϵ-LDP) variant to provide stronger privacy guarantees, the utility suffers when a large amount of noise is added (e.g. the recall drops below 85% for any ϵ value lower than 3). Considering the same system model and kinship coefficient metric, we propose an efficient framework that is able to identify higher degree of kinship relationships with lower privacy risks.

In this paper, we propose the Incremental Update Kinship Identification (INK) framework in which all the researchers generate the metadata and monitor the kinship relationships of each pair of individuals in the federated setting iteratively. We propose to make use of auxiliary information (e.g. simulation distributions of relatedness coefficients) to decide the cutoff values for classifying individuals belonging to different kinship categories. We define the informative score of each SNP based on the impact it has on the relatedness coefficients and select SNPs according to the ranks. In our work, we aim to accurately compute the kinship coefficients between all the samples in federated datasets and show that our approach has lower privacy risks than the previous works.

Since our framework is based on multiple components, we evaluate different methods that are built on a combination of multiple components. We compare them with the original centralized approach [14] and our previous approach [13]. We noticed that the usage of auxiliary information to determine cut-off points has a higher impact than the other framework components. Furthermore, we also explore the change in utility when we use different number of SNPs. Compared to methods that outsource metadata to the server, our method achieves higher kinship identification correctness while requiring fewer SNPs to be shared with the server, thus minimizing the privacy risk.

Kinship Inference

There exist a few metrics to measure the relatedness between two samples such as KING coefficient [15], Identity By Descent estimation method [16], Graphical Representation of Relationship errors [17], and KIND [14]. In this work, we use KING coefficient as the kinship metric because of its high accuracy and simplicity in use. Assuming that SNPs are biallelic, we compute the KING kinship coefficient between two samples i and j as follows:

$$\phi_{i,j} = \frac{2n_{11} - 4(n_{02} + n_{20}) - n_{*1} + n_{1*}}{4n_{1*}}, \tag{1}$$

where n_{11} represents the number of SNPs in which both samples are heterozygous, n_{02} in which sample i is homozygous dominant and sample j is homozygous recessive, n_{20} refers to the opposite case (i.e., i is homozygous recessive and j is homozygous dominant), n_{1*} and n_{*1} in which samples i and j are heterozygous, respectively.

Based on KING, a kinship coefficient greater than different thresholds implies different degrees of kinship relatedness.

Environment and Threat Model

System model. We refer to the parties that want to perform the collaborative study as researchers, and the third party that has the computation power as the server. The goal of each researcher is to identify the kinship relationships between samples across federated datasets while guaranteeing the privacy of the dataset participants (samples). The researchers need to share with the server some partial metadata which are locally generated from the original dataset. Then, the server performs the necessary computations to identify all the related samples across federated datasets. Based on the study that the researchers want to conduct (i.e., whether they need to remove related samples or keep only the family members), the server sends back the IDs of the samples that need to be filtered out to each researcher.

Threat model. There are several known privacy attacks in the field of genomic data science, such as membership inference attacks [18, 19], attribute inference attack [20], and reconstruction attacks [21]. In most cases, the adversaries are assumed to know the target's full or partial genomic sequences and exploit side information to increase the power of the attacks. In this work, we focus on general privacy risks without specifying particular attacks, and the goal is to limit the privacy risk due to sharing metadata iteratively.

We assume an honest-but-curious server and legitimate researchers. Suppose the server has a target and the partial genome sequence of the target. The server follows the protocol but tries to obtain more information, i.e., the genome sequences of the target. In order to conduct further attacks, such as membership inference attacks and reconstruction attacks, the server tries to match the target's sequence with the shared metadata from both researchers. This is achieved if (a) at each iteration, the server can match the unordered SNP sequence shared by the researchers with the target's

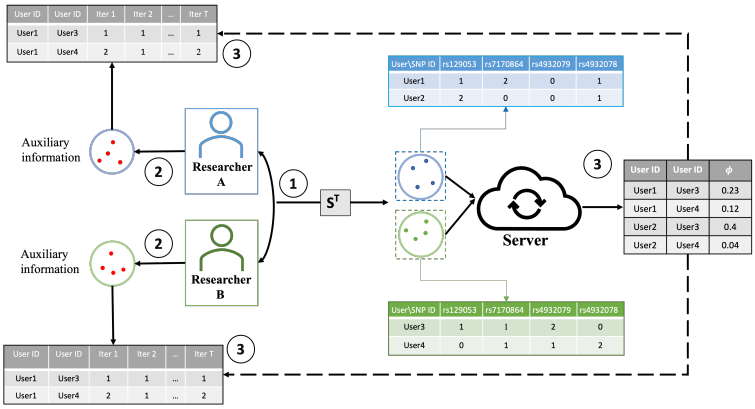

Figure 1: Workflow overview of the proposed framework. (1) Synchronization: two researchers initially decide whether they will continue the process; if continue, they decide on the set of SNPs that is used to generate the metadata which is later sent to the server. (2) Initialization: each researcher utilizes the auxiliary information (e.g. simulated samples) and extracts the basic statistics of the KING coefficients distributions. (3) Outsourcing: the server calculates and returns the pairwise KING coefficients to both researchers. Each researcher updates locally the kinship relationships.

SNP sequence and (b) the server is able to find the linkage among the SNPs shared across the iteration process. The privacy risk, therefore, depends on the number of shared SNPs, the number of iterations, and the correlations between SNPs.

Methods

In this section, we describe the proposed mechanism to identify kinship relationships in the federated setting that achieves high kinship identification accuracy and is robust even when a small set of SNPs is provided. We initially present the proposed Incremental Update Kinship Identification (INK) framework. Then, we describe the three components, including exploiting auxiliary information, incremental process, and informative score, which pave the way to address the drawbacks of existing methods (discussed in the previous section).

Proposed Framework. In the following sections, for the ease of simplicity, we have considered two researchers, but our framework can be easily extended to multiple researchers. The algorithm of the proposed framework is shown in Algorithm 1 and the overall workflow is illustrated in Figure 1. Similar to the previous method [13], each researcher sends some metadata to the server to facilitate computation of the KING coefficients. At each iteration, there are three stages which include: synchronization stage, initialization stage, and outsourcing stage. In the synchronization stage, the researchers mutually decide on the set of SNPs which are used to generate the metadata. The selected set of SNPs (i.e. SNP IDs) is not revealed to the server as part of the metadata. In the initialization stage, each researcher generates synthetic relatives and unrelated individuals (i.e., simulated set) following Mendel's Laws. Each researcher calculates the KING coefficients on the simulated set and extracts the aggregate statistics for each kinship relationship group, i.e., $([\min, \max], mean)_k$ (line 8 to 11). Note that the researchers only synchronize to decide the metadata.

In the outsourcing stage, both researchers send the metadata to the server. Since the metadata does not contain the SNP IDs, the server cannot simply construct the SNP sequence of the samples that are included in the metadata. The server calculates the KING coefficients with the pooled metadata and returns each pair's coefficients to both researchers. Each researcher calculates the KING coefficients on all the shared SNPs via Eq 2. Then, comparing the results with the aggregate statistics, each researcher obtains the kinship relationship for each pair of individuals. Finally, each researcher updates the kinship relationship based on the previous results using a weighted voting strategy (line 14 to 19). Given a list of kinship relationship predictions from T iterations, $k^1, k^2, \cdots, k^T$, the weighted kinship relationship k is calculated as $\sum_1^T w^t * k^t$, where w^t is the weight associates with k^t, and $w_t = \frac{\frac{1}{T+1-t}}{\sum_t^T \frac{1}{T+1-t}}$. Each researcher then takes the majority weighted vote, $k^* = \text{argmax}\, k$, and sets it as the kinship relationship between the samples. In the next synchronization phase, the researchers also need to decide if they want to continue. If the kinship

Algorithm 1 Incremental Update Kinship Identification (INK)

Input: Individuals data D_A and D_B from researcher A and B
Output: Kinship relatedness of the individuals in D_A and D_B

Initialize $S = \emptyset$
$T = 1$
while stop condition is not satisfied **do**
    Sync Stage:
        Both researchers, A and B, decide S^T to be shared to the server
        $S = S \cup S^T$
    Init Stage:
        for r = A and B **do**
            Generate simulated unrelated individuals and relatives, Sim_D^r, of D_r
            Calculate KING coefficients with Eq. 1 of Sim_D^r on S
            Extract $([\min, \max], mean)_k^r$ of the coefficients of the k-degree relatives
                in Sim_D^r where $k \in \{1, 2, 3, unrelated\}$
        end for
    Outsourcing Stage:
        A and B share S^T of D_A and D_B to the server, respectively
        Sever computes the KING coefficients and return $\{i \in D_A, j \in D_B | (i, j, \phi_{i,j}^T)\}$ to A and B
        for r = A and B **do**
            Compute $\phi_{i,j}$ based on Eq. 2 and $\phi_{i,j}^T$
            Obtain the kinship relationship
                $k^T = \underset{k}{\operatorname{argmin}} |\phi_R - mean_k^r|$, where $\phi_R \in [min, max]_k^r$ and $k \in [1, 2, 3, unrelated]$
            Update the kinship relationship
                $k^* = \underset{k}{\operatorname{argmax}} \sum_1^T w^t * k^t$, where $w_t = \frac{\frac{1}{T+1-t}}{\sum_t^T \frac{1}{T+1-t}}$
        end for
    $T = T + 1$
end while

relationships for all the individuals remain the same, both researchers choose to end the iteration.

Exploiting Auxiliary Information. Determining the kinship relaq1ture of the population [22]. A large set of SNPs can help to increase the accuracy and robustness of the relatedness measure-based estimators (e.g., KING). In fact, the distributions of the KING coefficient for different kinship categories (i.e., relationship degrees) are based on 20K SNPs [15]. However, using a large set of SNPs is not applicable in the federated setting for two reasons: (a) a large number of overlapping SNPs may not be feasible for all the participating parties; and (b) sharing a large number of SNPs can significantly increase the privacy risks.

To address this issue, we propose to use the auxiliary information, such as simulated distributions of KING coefficients and publicly available genomic datasets which contain real relatives, to determine the cutoff values (i.e., thresholds) while classifying KING coefficients as belonging to these kinship categories.

In the proposed framework, each researcher r creates a simulated set, denoted as Sim^r, that contains artificially generated relatives (up to third-degree relatives) and unrelated individuals of the individuals in their dataset. Given a set of SNPs and a pair of real individuals (i, j), the researcher initially calculates the KING coefficients of the simulated kinship categories (k) and obtains the descriptive statistics, i.e., $\{minimal, maximal, mean\}_k$, of the corresponding coefficients distributions. Then, the researcher classifies the kinship relationship of the pair of real individuals based on the KING coefficients ($\phi_{i,j}$) and the distance to the simulated distributions. If the $\phi_{i,j} \in [min, max]_k$, it is classified as a k-th degree kinship category. If the coefficient falls into multiple categories, the kinship relation is determined by the distance to the $mean$, and the closest category is chosen, i.e., $k = \underset{k}{\operatorname{argmin}} |\phi_R - mean_k|$ where $\phi_R \in [min, max]_k$

and $k \in [1, 2, 3, unrelated]$.

Incremental Process. Besides the aforementioned factors, the relatedness measure-based estimators suffer from the inherent inaccuracies due to the inconsistencies between genetic and pedigree relatedness, which leads to overlapping distributions of coefficients for different kinship categories [23, 15]. This, however, holds even with a large number of SNPs and/or the auxiliary information (e.g., simulated distributions). To address the issue, we propose an incremental process. Our proposed process starts with a small number of SNPs, and then gradually expands to include more SNPs. This approach is both robust and efficient, and it can handle large datasets.

Instead of outsourcing metadata all at once, each researcher breaks down the metadata and outsources a small subset iteratively. Similar to the ensemble technique, researchers combine the results from multiple iterations to improve the robustness of the relatedness-based approach. Moreover, with the proposed outsourcing framework, each researcher is able to recover the results as computed on the whole metadata based on the results returned at each iteration. Assume that at iterations t and $t+p$, researchers share metadata M^t and M^{t+p} to the server. Note that each metadata includes a small set of SNPs, denoted as S^t and S^{t+p} respectively, and $S^t \cap S^{t+p} = \emptyset$. Server returns the KING coefficients computed on S^t and S^{t+p} to both researchers. Denote ϕ, ϕ^t, and ϕ^{t+p} as the KING coefficients computed on $S^t \cup S^{t+p}$, S^t, and S^{t+p} respectively. Then, both researchers obtain ϕ as:

$$\phi = \frac{n_{1*}^t \phi^t + n_{1*}^{t+p} \phi^{t+p}}{n_{1*}^t + n_{1*}^{t+p}} \tag{2}$$

Note that this is true since

$$\begin{aligned}
\phi =& \frac{2n_{11} - 4(n_{02} + n_{20}) - n_{*1} + n_{1*}}{4n_{1*}} \\
=& \frac{2\left(n_{11}^t + n_{11}^{t+p}\right) - 4\left(\left(n_{02}^t + n_{02}^{t+p}\right) + \left(n_{20}^t + n_{20}^{t+p}\right)\right) - \left(n_{*1}^t + n_{*1}^{t+p}\right) + \left(n_{1*}^t + n_{1*}^{t+p}\right)}{4\left(n_{1*}^t + n_{1*}^{t+p}\right)} \\
=& \frac{n_{1*}^t \phi^t + n_{1*}^{t+p} \phi^{t+p}}{\left(n_{1*}^t + n_{1*}^{t+p}\right)}
\end{aligned} \tag{3}$$

Informative Score. Given SNP sequences from two individuals, the kinship relatedness based on the KING coefficient is determined by the pairwise combinations of SNPs (the total number of SNPs in which individuals are heterozygous, homozygous dominant, and homozygous recessive). When sharing of the whole SNP sequence is infeasible (as in the federated setting), the SNPs that are selected to compute the KING coefficients are important in order to achieve high accuracy in an efficient way. For instance, outsourcing a set of SNPs that have small impact on the KING coefficient (i.e., SNPs for which both individuals are homozygous recessive or homozygous dominant) is undesirable. In this section, we propose informative score (IS) to measure the impact of each SNP on KING coefficients calculation. In the federated setting, the IS of each SNP is not directly available since the individuals are separated in two different datasets. Therefore, we utilize the simulated dataset, as described in the previous section, to calculate the informative score for each SNP. Formally, we define the informative score of SNP h as:

$$IS(h) = \frac{1}{|\sigma|} \sum_{(i,i) \in \sigma} \left| \log \frac{\phi_{i,j}^{S}}{\phi_{i,j}^{S_{-h}}} \right| \tag{4}$$

where σ denotes all the possible combinations of samples in the simulation set and S_{-h} denotes the SNPs set without h (i.e., $S_{-h} = S \setminus \{h\}$).

During the iteration process, researchers first outsource the SNPs with high informative scores. When a set of SNPs have the same informative scores, a random selection is applied to break ties.

Privacy Analysis

As mentioned before, in this work, we focus on general privacy risks without specifying one particular attack, and the goal is to limit the privacy risk due to sharing metadata iteratively. The server is able to infer the sensitive information

from the metadata if (a) at each iteration, the server can match the unordered SNP sequence shared by the researchers with a target's SNP sequence (i.e., a target individual, whose membership to a research dataset may be sensitive) and (b) the server is able to find the linkage among the SNPs shared across the iteration process.

To alleviate the privacy risk, we present two mitigation techniques. On the one hand, researchers select SNPs from different chromosomes for each iteration to remove the linkage among SNPs in different iterations. Researchers also remove the user IDs and shuffle the samples in each iteration. On the other hand, during each iteration, each researcher generates several synthetic SNPs and shares them as part of the metadata with the server. With carefully tuned parameters, the synthetic SNPs are indistinguishable from the real ones [24]. To remove the noise introduced by the synthetic SNPs, researchers initially create two different sets of synthetic SNPs, named S^{F_1} and S^{F_2}, then share three batches of metadata, $S^{F^1} \cup S^R$, $S^{F_2} \cup S^R$, and $S^{F_1} \cup S^{F_2}$ respectively, with the server and obtain the KING coefficients, ϕ^{F_1R}, ϕ^{F_2R}, and $\phi^{F_1F_2}$. Based on Eq. 2, each researcher recovers the KING coefficients calculated on the real SNPs set R, ϕ^R, with

$$\phi^R = \frac{n_{1*}^{F_1R}\phi^{F_1R} + n_{1*}^{F_2R}\phi^{F_2R} - n_{1*}^{F_1F_2}\phi^{F_1F_2}}{2n_{1*}^R}$$

Note that each researcher generates the synthetic SNPs independently.

Experiments

As discussed in the previous section, our proposed framework is based on three components. We propose several methods based on the components (see Table 1) and their combinations as follows:

1. Increment: Iteratively outsource the SNPs and update the kinship relationships based on the predefined thresholds
2. Increment+Sim: Iteratively outsource the SNPs and update the kinship relationships based on the simulated thresholds
3. Increment+Informative: Iteratively outsource the SNPs based on the informative scores, and update the kinship relationships based on the predefined thresholds
4. Increment+Inform+Sim: Iteratively outsource the SNPs based on the informative scores, and update the kinship relationships based on the simulated thresholds

	Increment		Sim		Inform		All	
	outsource	update	outsource	update	outsource	update	outsource	update
Increment	⋆	⋆	†	†	•	⋆		
Sim	†	†	—	—	—	—	•	†
Inform	•	⋆	—	—	—	—		

Note: Predefined threshold (⋆) ; Simulated threshold (†) ; Informative scores (•) ; Not applicable (—)

Table 1: Summary of components combinations

We compare the proposed methods with two baselines: centralized approach and privacy-preserving kinship identification (PPKI) [13]. The centralized approach assumes all the data is stored in one local node, and the computation is carried out locally. PPKI assumes there are two researchers and utilize an ϵ-LDP variant for preserving privacy. ϵ is the privacy parameter which controls the trade-off between privacy and accuracy. Following the implementation of the PPKI, we pick the parameters that result in high utility, and therefore, we set ϵ to be 5.

To evaluate the proposed methods, we use real genomic data from OpenSNP [25] and randomly sample a subset that contains 200 unrelated individuals and 3000 SNPs. Following Mendel's law, we generate 200 pairs of synthetic relatives for each degree of kinship relationship (up to third-degree) of the individuals in the subset.

The performance of the proposed methods in comparison with the two baselines is evaluated in terms of utility (measured by the correctness of identifying kinship relationships) and the relative privacy loss (measured by the number of SNPs required to achieve a certain correctness level). The method that achieves higher kinship identification correctness and lower relative privacy loss is preferred.

Kinship Identification Correctness. We use accuracy rate and Area Under the Receiver Operating Characteristic Curve (AUROC) scores to measure the correctness of kinship relationships identification.

Relative Privacy Loss. As discussed in the previous section, we use the number of outsourced SNPs to measure the potential privacy risk in the federated setting. Denote the number of SNPs required by baseline q as $|S|^q$ and the kinship identification correctness achieved by this baseline with $|S|^q$ SNPs as c-correctness. We then define the relative privacy loss ($RPL(c)$) of approach p to the baseline q as

$$RPL(c)_{p,q} = \frac{|S|^p - |S|^q}{|S|^q} \tag{5}$$

where $|S|^p$ is the the minimal number of SNPs required by q to achieve c-correctness. In the following experiments, we use the centralized method, i.e., all the data is stored in a single node and no privacy risk introduced, as the baseline q. As per the above definition, a negative value is always preferred since it means a privacy gain compared to the centralized approach.

Results

Our objective is to identify the kinship correlation of individuals from different data sites. Here, we consider two practical tasks (a) identifying the degree of kinship relatedness (we consider first, second, third-degree relatives and unrelated samples) and (b) identifying relatives and unrelated individuals. We conduct each experiment 10 times and report the average of the results.

Utility. The utility of identifying relatives is shown in Table 2. For the proposed methods, we fix the size of the metadata (the number of SNPs) in each iteration at 100 and continue the iteration until all the SNPs are shared with the server. For the two baselines, we calculate the KING coefficients using all the SNPs. From Table 2, we observe that under task (a), all the proposed methods outperform PPKI [13]. *Simulation set* has a more significant impact than the other two components. Increment+Sim and Increment+Inform+Sim achieve higher accuracy rates and AUPRC scores compared to the other methods. Under task (b), we notice that *Simulation set* has a minor contribution. This is due to the low expressibility of the basic statistics including minimum, maximum, and mean. Samples containing different kinship categories can lead to a complicated KING coefficient distribution, which basic statistics cannot capture. To address this issue, we can incorporate the misclassification rate to determine the cutoff values [23]. However, the utility is still improved solely by the *Incremental Process* and the weighted voting strategy.

In Figure 2, we further explore the change in utility for all the kinship relationships with different numbers of SNPs. The correctness of identifying kinship relatedness increases when more SNPs are provided. However, to identify higher degree kinship relationship, more SNPs are required. While no single method dominates the rest, *Incremental process* and *Simulation set* show a greater improvement in terms of utility.

	Task (a)		Task (b)	
Methods	Accuracy	AUROC	Accuracy	AUROC
Increment	0.803	0.870	0.908	0.868
Increment+Sim	0.821	0.881	0.814	0.858
Increment+Informative	0.800	0.869	0.877	0.808
Increment+Inform+Sim	0.816	0.877	0.900	0.800
Centralized	0.807	0.873	0.906	0.866
PPKI [13]($\epsilon = 5$)	0.799	0.866	0.903	0.866

Table 2: Kinship identification correctness. The best performance is marked in light blue font, and the second best is marked in blue font.

Relative Privacy Loss. Based on the accuracy rate and AUPRC scores of the centralized setting (which is considered as the main baseline in this experiment), we compute the relative privacy loss of the proposed methods and PPKI [13], and present them in Table 3. Note that > 0 suggests that the relative privacy loss is undefined since the number of the required SNPs is larger than the dataset size. Increment+Sim achieves the best and the second best RPL results in terms of AUROC for tasks (a) and (b), respectively. The obtained results show that Increment+Sim require 24% fewer SNPs in order to achieve the same kinship identification correctness as the baseline.

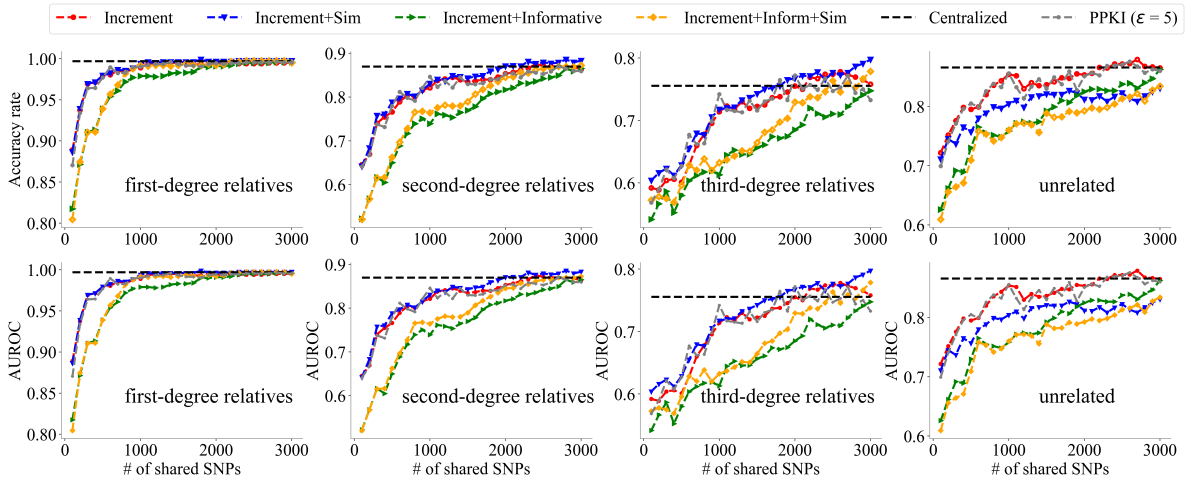

Figure 2: Utility for different degrees of relatedness with different number of SNPs. Centralized is evaluated with all the 3000 SNPs.

	Task (a)		Task (b)	
Methods	RPL(acc)	RPL(auc)	RPL(acc)	RPL(auc)
Increment	> 0	> 0	-0.19	-0.09
Increment+Sim	-0.25	-0.24	> 0	-0.09
Increment+Informative	> 0	> 0	> 0	> 0
Increment+Inform+Sim	-0.03	> 0	> 0	> 0
PPKI [13] ($\epsilon = 5$)	-0.05	> 0	-0.24	-0.19

Table 3: Relative privacy loss. The best performance is marked in light blue font, and the second best is marked in blue font.

Efficiency and Utility Trade-off. The proposed framework assumes that the researchers need to synchronize and maintain the samples' kinship relatedness once the server returns the results at each iteration. Hence, the computation and communication overhead increases linearly with the number of iterations. Researchers can stop the iteration process early if the goal is to identify the lower degrees of relatives, as shown in Figure 2 (e.g., only a small number of SNPs, 200-400, enables to achieve a high accuracy rate of identifying first-degree relatives). In order to identify a higher degree relatives and reduce the computation and communication overhead, researchers can increase the number of shared SNPs (m) at each iteration. In Table 4, we show the change in utility when considering different numbers of shared SNPs in an iteration. From Table 4, the communication and computation overhead can be reduced drastically with a low impact on the utility. For instance, when m increases from 100 to 600, the number of iterations decreases from 30 to 5, and the AUPRC score of Increment+Sim slightly decreases from 0.881 to 0.879 for task (a).

Conclusion

The volume and pace at which genomic data is collected and used to train models keeps on increasing enormously [26, 27, 28, 29]. Since sending out the sensitive individual genomic data is prohibited by regulations like HIPAA,

		Task (a)				Task (b)			
	Methods	$m = 100$	$m = 200$	$m = 400$	$m = 600$	$m = 100$	$m = 200$	$m = 400$	$m = 600$
Accuracy	Increment	0.806	0.805	0.800	0.802	0.912	0.911	0.910	0.908
	Increment+Sim	0.823	0.822	0.817	0.819	0.832	0.829	0.828	0.825
	Increment+Informative	0.800	0.791	0.787	0.790	0.879	0.884	0.886	0.889
	Increment+Inform+Sim	0.816	0.813	0.803	0.804	0.900	0.899	0.893	0.899
AUROC	Increment	0.870	0.870	0.867	0.869	0.868	0.866	0.863	0.861
	Increment+Sim	0.881	0.881	0.877	0.879	0.858	0.858	0.858	0.858
	Increment+Informative	0.869	0.864	0.860	0.862	0.808	0.820	0.833	0.848
	Increment+Inform+Sim	0.877	0.876	0.869	0.870	0.800	0.798	0.785	0.803

Table 4: Utility with varying numbers of shared SNPs in each iteration. Results are based on 3000 SNPs in total.

studies that require population-wise correlation cannot be done in a centralized way and a privacy-preserving method to estimate the kinship relationship is necessary to balance privacy concerns and research utility.

In this work, we propose the Incremental Update Kinship Identification (INK), which shows considerable improvements to the prior Privacy-preserving Kinship Identification (PPKI) method [13]. Our proposed kinship framework is the state-of-the-art method for estimating kinship relationships in a federated setting while preserving privacy. By sharing limited number of SNPs, INK can achieve a good accuracy in estimating the degree of relatedness. This allows us to estimate kinship relationships while preserving privacy by limiting the amount of information shared. Moreover, the flexibility of combining three novel components (Incremental process, Informative score, and Auxiliary information) in INK framework allow the flexibility to deal with heterogeneous datasets. Collaborators can choose the best performing combination of these components according to their required kinship estimation accuracy.

Our proposed method is lightweight, easy to implement on the researcher's side, and does not require a completely trusted server. However, the implementation of real-world applications needs to handle problems including client selection, dropout, and heterogeneous data, which can be explored in future work. Furthermore, the experimental setting assumes that the sample sizes in each node are equal. In general, this is not true. Note that in a federated network, the node with smaller sample size will be more vulnerable to others when it comes to privacy protection. With this in mind, our future work will focus on adopting secure multi-party computation (secure-MPC) protocols to the INK framework, such as ABY^3 [30], Falcon [31], Function Secret Sharing (FSS) [32], SPDZ [33, 34], etc, and evaluating different settings including having multiple parties.

Acknowledgments

XJ is CPRIT Scholar in Cancer Research (RR180012), and was supported in part by Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, the National Institutes of Health (NIH) under award R01AG066749, R01LM013712, and U01TR002062, and the National Science Foundation (NSF) #2124789. EA was partly supported by the National Library of Medicine of the NIH under award R01LM013429 and by the NSF under grant numbers 2141622, 2050410, 2200255, and OAC-2112606. JV was partly supported by the NIH under award R35GM134927, and a research gift received from Cisco University Research.

References

1. Birney E, Vamathevan J, Goodhand P. Genomics in healthcare: GA4GH looks to 2022. BioRxiv. 2017:203554.
2. Stark Z, Dolman L, Manolio TA, Ozenberger B, Hill SL, Caulfied MJ, et al. Integrating genomics into healthcare: a global responsibility. The American Journal of Human Genetics. 2019;104(1):13-20.
3. Lee SSJ, Borgelt E. Protecting posted genes: Social networking and the limits of GINA. The American Journal of Bioethics. 2014;14(11):32-44.
4. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nature genetics. 2020;52(7):646-54.
5. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circulation: Cardiovascular Quality and Outcomes. 2019;12(7):e005122.
6. Al Aziz MM, Anjum MM, Mohammed N, Jiang X. Generalized Genomic Data Sharing for Differentially Private Federated Learning. Journal of Biomedical Informatics. 2022:104113.
7. Singh AP, Zafer S, Pe'er I. MetaSeq: privacy preserving meta-analysis of sequencing-based association studies. In: Biocomputing 2013. World Scientific; 2013. p. 356-67.
8. Xie W, Kantarcioglu M, Bush WS, Crawford D, Denny JC, Heatherly R, et al. SecureMA: protecting participant privacy in genetic association meta-analysis. Bioinformatics. 2014;30(23):3334-41.
9. Sadat MN, Al Aziz MM, Mohammed N, Chen F, Wang S, Jiang X. SAFETY: Secure gwAs in Federated Environment Through a hYbrid solution with Intel SGX and Homomorphic Encryption. arXiv: 170302577. 2017.
10. Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nature biotechnology. 2018;36(6):547-51.
11. Jha S, Kruger L, Shmatikov V. Towards practical privacy for genomic computation. In: 2008 IEEE Symposium on Security and Privacy (sp 2008). IEEE; 2008. p. 216-30.

12. Zhu D, Zhu H, Wang X, Lu R, Feng D. Efficient and Privacy-preserving Similar Patients Query Scheme over Outsourced Genomic Data. IEEE Transactions on Cloud Computing. 2021.
13. Dervishi L, Wang X, Li W, Halimi A, Vaidya J, Jiang X, et al. Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy. In Proc of the 2022 AMIA Annual Symposium. 2022.
14. Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. The American Journal of Human Genetics. 2008;82(2):352-65.
15. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
16. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics. 2011;88(2):173-82.
17. Abecasis GR, Cherny SS, Cookson W, Cardon LR. GRR: graphical representation of relationship errors. Bioinformatics. 2001;17(8):742-3.
18. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS genetics. 2008;4(8):e1000167.
19. Wang R, Li YF, Wang X, Tang H, Zhou X. Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of ACM SIGSAC CCS; 2009. p. 534-44.
20. Humbert M, Ayday E, Hubaux JP, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: Proceedings of ACM SIGSAC CCS; 2013. p. 1141-52.
21. Mizas C, Sirakoulis GC, Mardiris V, Karafyllidis I, Glykos N, Sandaltzopoulos R. Reconstruction of DNA sequences using genetic algorithms and cellular automata: Towards mutation prediction? Biosystems. 2008;92(1):61-8.
22. Städele V, Vigilant L. Strategies for determining kinship in wild populations using genetic data. Ecology and Evolution. 2016;6(17):6107-20.
23. Blouin M, Parsons M, Lacaille V, Lotz S. Use of microsatellite loci to classify individuals by relatedness. Molecular ecology. 1996;5(3):393-401.
24. Wang X, Jiang X, Vaidya J. Efficient verification for outsourced genome-wide association studies. Journal of Biomedical Informatics. 2021;117:103714.
25. Greshake B, Bayer PE, Rausch H, Reda J. OpenSNP-a crowdsourced web resource for personal genomics. PLOS ONE. 2014;9(3):e89204.
26. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. Responsible, practical genomic data sharing that accelerates research. Nature Reviews Genetics. 2020;21(10):615-29.
27. Liu H, Guo G. Opportunities and challenges of big data for the social sciences: The case of genomic data. Social science research. 2016;59:13-22.
28. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods. 2019;166:4-21.
29. Kashyap H, Ahmed HA, Hoque N, Roy S, Bhattacharyya DK. Big data analytics in bioinformatics: A machine learning perspective. arXiv preprint arXiv:150605101. 2015.
30. Mohassel P, Rindal P. ABY3: A mixed protocol framework for machine learning. In: Proceedings of ACM SIGSAC CCS; 2018. p. 35-52.
31. Wagh S, Tople S, Benhamouda F, Kushilevitz E, Mittal P, Rabin T. Falcon: Honest-majority maliciously secure framework for private deep learning. arXiv preprint arXiv:200402229. 2020.
32. Boyle E, Gilboa N, Ishai Y. Function secret sharing. In: Annual international conference on the theory and applications of cryptographic techniques. Springer; 2015. p. 337-67.
33. Damgård I, Pastro V, Smart N, Zakarias S. Multiparty computation from somewhat homomorphic encryption. In: Annual Cryptology Conference. Springer; 2012. p. 643-62.
34. Keller M. MP-SPDZ: A versatile framework for multi-party computation. In: Proceedings of ACM SIGSAC CCS; 2020. p. 1575-90.