Poster: Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy

Leonard Dervishi
Case Western Reserve University
lxd244@case.edu

Xinyue Wang
Rutgers University
xw273@scarletmail.rutgers.edu

Wentao Li
University of Texas Health School of Biomedical Informatics
wentao.li@uth.tmc.edu

Anisa Halimi
IBM Research Europe
anisa.halimi@ibm.com

Jaideep Vaidya
Rutgers University
jsvaidya@business.rutgers.edu

Xiaoqian Jiang
University of Texas Health School of Biomedical Informatics
xiaoqian.jiang@uth.tmc.edu

Erman Ayday
Case Western Reserve University
exa208@case.edu

Abstract

With the reduction of sequencing costs and the pervasiveness of computing devices, genomic data collection is continually growing. However, data collection is highly fragmented and the data is still siloed across different repositories. Analyzing all of this data would be transformative for genomics research. However, the data is sensitive, and therefore cannot be easily centralized. Furthermore, there may be correlations in the data, which if not detected, can impact the analysis. In this paper, we take the first step towards identifying correlated records across multiple data repositories in a privacy-preserving manner. The proposed framework, based on random shuffling, synthetic record generation, and local differential privacy, allows a trade-off of accuracy and computational efficiency. An extensive evaluation on real genomic data from the OpenSNP dataset shows that the proposed solution is efficient and effective.

I. Reference

Dervishi L, Wang X, Li W, Halimi A, Vaidya J, Jiang X, Ayday E. "Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy" in AMIA 2022 Annual Symposium, Washington, DC, USA, 2022.

DOI: Currently not available. The link to the pre-print version: https://arxiv.org/pdf/2203.05664.pdf

II. Acknowledgements

The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy

Leonard Dervishi, Xinyue Wang, Wentao Li, Anisa Halimi,
Jaideep Vaidya, Xiaoqian Jiang, Erman Ayday

Scan the QR code to access the full paper

Our framework identifies correlated records across multiple genomic data repositories **in a privacy preserving manner.**

1. Introduction

- Collaborative research produces more accurate outcomes and powerful statistics
- Sharing of genomic data has privacy implications
- It is crucial to perform sample relatedness as part of quality control

2. Methods

A. Researchers coordinate to decide on the set of SNPs that will be shared with the server
B. The generation of the metadata M^i from the original dataset D^i
C. Each researcher R^i sends their prepared metadata M^i to the server, which computes the pairwise kinship coefficients among all samples

3. Results

- 95% kinship accuracy when the shared number of SNPs is m=500 and the privacy coefficient ϵ=5

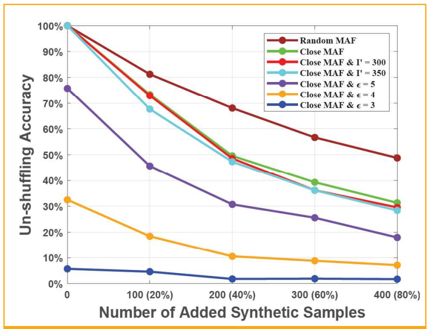


Figure 2. The server's accuracy in un-shuffling the shared SNPs in the metadata for different scenarios.

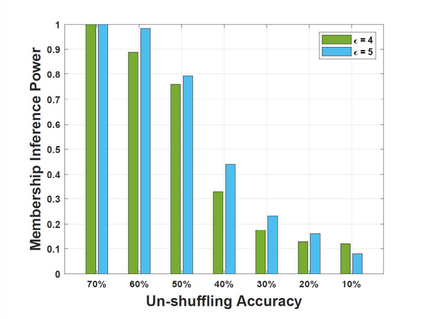


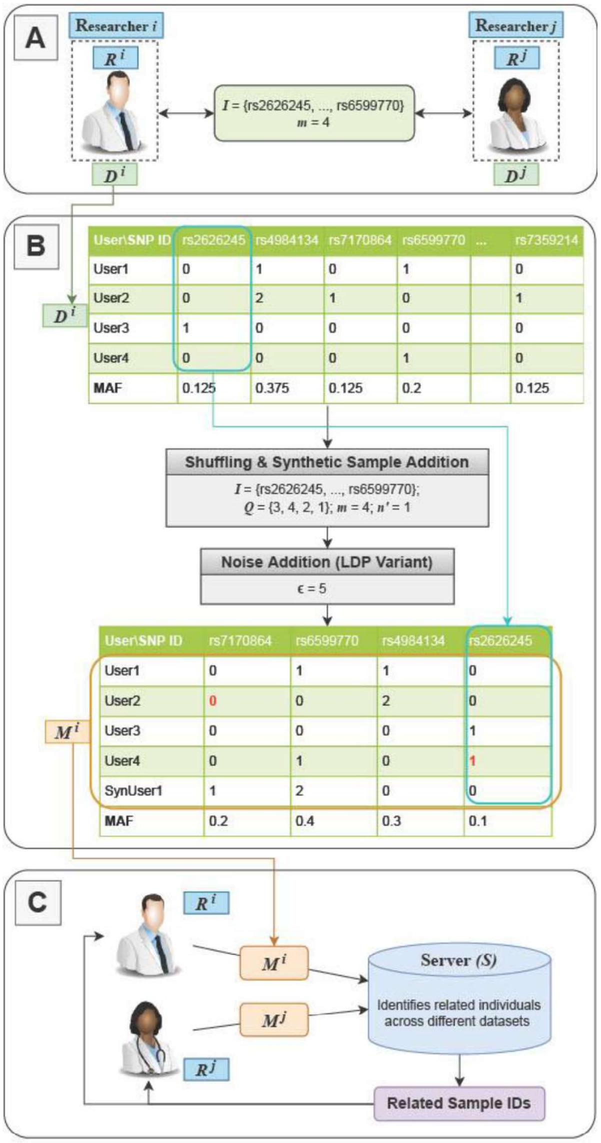
Figure 3. Power of membership inference attack for different ϵ values considering various un-shuffling accuracy results

4. Privacy Analysis

- Privacy risk consists of the server un-shuffling the shared SNPs
- We use a greedy algorithm to Iteratively match SNPs based on MAFs and SNP correlations
- Power analysis based on hamming distance

5. Conclusion

- Efficient and effective privacy-preserving technique to identify correlated samples across federated datasets
- Fine tuning parameters to control the trade-off between the accuracy, privacy and computational load



User\SNP ID	rs2626245	rs4984134	rs7170864	rs6599770	...	rs7359214
User1	0	1	0	1		0
User2	0	2	1	0		1
User3	1	0	0	0		0
User4	0	0	0	1		0
MAF	0.125	0.375	0.125	0.2		0.125

Shuffling & Synthetic Sample Addition

I = {rs2626245, ..., rs6599770};
Q = {3, 4, 2, 1}; m = 4; n' = 1

Noise Addition (LDP Variant)

ϵ = 5

User\SNP ID	rs7170864	rs6599770	rs4984134	rs2626245
User1	0	1	1	0
User2	0	0	2	0
User3	0	0	0	1
User4	0	1	0	1
SynUser1	1	2	0	0
MAF	0.2	0.4	0.3	0.1

Identifies related individuals across different datasets

Related Sample IDs

Figure 1. General overview of the proposed framework.

Acknowledgements

The work was partly supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013429 and by the National Science Foundation (NSF) under grant numbers 2141622, 2050410, 2200255, and OAC-2112606.

References

1. *Manichaikul, Ani, et al. "Robust relationship inference in genome-wide association studies." Bioinformatics 26.22 (2010): 2867-2873.*
2. *Cho, Hyunghoon, David J. Wu, and Bonnie Berger. "Secure genome-wide association analysis using multiparty computation." Nature biotechnology 36.6 (2018): 547-551.*
3. *Halimi, Anisa, et al. "Privacy-preserving and efficient verification of the outcome in genome-wide association studies." arXiv preprint arXiv:2101.08879 (2021).*