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

Leonard Dervishi
Case Western Reserve University
lx244@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


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.
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 $\epsilon=5$

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

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-2112616.

References