Poster: Reproducibility-Oriented and Privacy-Preserving Genomic Dataset Sharing

Yuzhou Jiang
Case Western Reserve University
yxj466@case.edu

Tianxi Ji
Texas Tech University
tiji@ttu.edu

Erman Ayday
Case Western Reserve University
exa208@case.edu

***Abstract*—The increasing pace in genomic research has brought a high demand for genomic datasets in recent years, yet few studies have released their datasets due to privacy concerns. This poses a problem while validating and reproducing the published results. In this work, in order to promote reproducibility of genome-related research, we propose a novel scheme for sharing genomic datasets under differential privacy. Our scheme shows great performance in terms of genome-wide association studies (GWAS) reproducibility, other data utility metrics, resistance against membership inference attacks, and running time. By constraining the privacy leakage, our mechanism is able to encourage the sharing of a genomic dataset along with the research results on it.**

I. Introduction

Development in genome sequencing has brought tremendous research opportunities in the genomic field in recent years. However, a major concern is that genomic research outcomes are hardly reproducible because other researchers can barely access the GWAS datasets that are used to produce the research outcomes due to privacy concerns. This prevents their peers from assessing the quality and correctness of the discoveries, and eventually impedes the development in the genomic area.

In this work, we target reproducibility of genomic research outcomes and propose a novel scheme that shares genomic datasets of point mutations on the DNA (i.e., Single Nucleotides Polymorphism - SNPs) under differential privacy. We focus on sharing SNP datasets, because such datasets are the most popular in biomedical research and GWAS. In the first step, the scheme generates a noisy copy of the genomic dataset by encoding the data entries as binary values and then XORing them with binary noise, that is calibrated and sampled with optimized time complexity, while considering the biology properties of the datasets. In the second step, the scheme alters the value distribution of each column in the generated copy to align with the privacy-preserving version (protected by the Laplace mechanism) of the distribution in the original dataset using optimal transport.

We implement our proposed scheme on two real genomic datasets from the OpenSNP project [2] and evaluate the scheme with regard to GWAS reproducibility. For comparison, we implemented two existing differentially private dataset sharing methods, i.e., DPSyn [6] and PrivBayes [7], which are both winning algorithms in the *NIST Differential Privacy Synthetic Data Challenge* [1] in 2018. The results prove that our scheme significantly outperforms the two existing methods in GWAS reproducibility and also other utility metrics. In addition, our scheme achieves better resistance against membership inference attacks, i.e., the Hamming distance attack [4] and machine-learning-based attacks,

II. System and Threat Model

We consider two parties involved in our system: a researcher and a verifier. The researcher has some research findings on a genomic dataset and wants to share the dataset for reproducibility, while a verifier wants to reproduce the research findings. The verifier can be the reviewers in a peer review publication or any other researcher who wants to reproduce the results for validation or comparison.

We assume the researcher is trusted. The researcher holds the original genomic dataset and will not expose it. While implementing the privacy mechanism, the researcher will not leak any information other than the shared genomic dataset, where such information includes the original dataset and all the intermediate data generated during privacy enhancement. We assume the verifier can be malicious and is curious about the original dataset. The malicious verifier, i.e., the attacker, may perform membership inference attacks (MIAs) to learn whether a target victim is a member of the noisy genomic dataset that is shared by the researcher.

We assume the attacker has access to the following knowledge from the researcher: (i) the shared dataset from the researcher and (ii) the trait/disease that the individuals in the dataset share. We consider two types of MIAs: the Hamming distance test and machine learning (ML)-based methods. The Hamming distance test (HDT) utilizes the Hamming distance between genomic records to measure the membership information of a given target. ML-based attacks are popular approaches used for membership inference tasks, and we consider two machine learning-based MIAs: random forest (RF) and support vector machine (SVM).

III. Methodology

The workflow of the proposed scheme is shown in Figure 1. There are two stages: data perturbation and utility restoration. In the data perturbation stage, we encode the genomic data into binary values, perturb the encoded dataset using the XOR mechanism [5], which is a state-of-the-art algorithm that releases binary data in matrix format while satisfying differential privacy, and then decode it back to the genomic domain. The

[1]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.

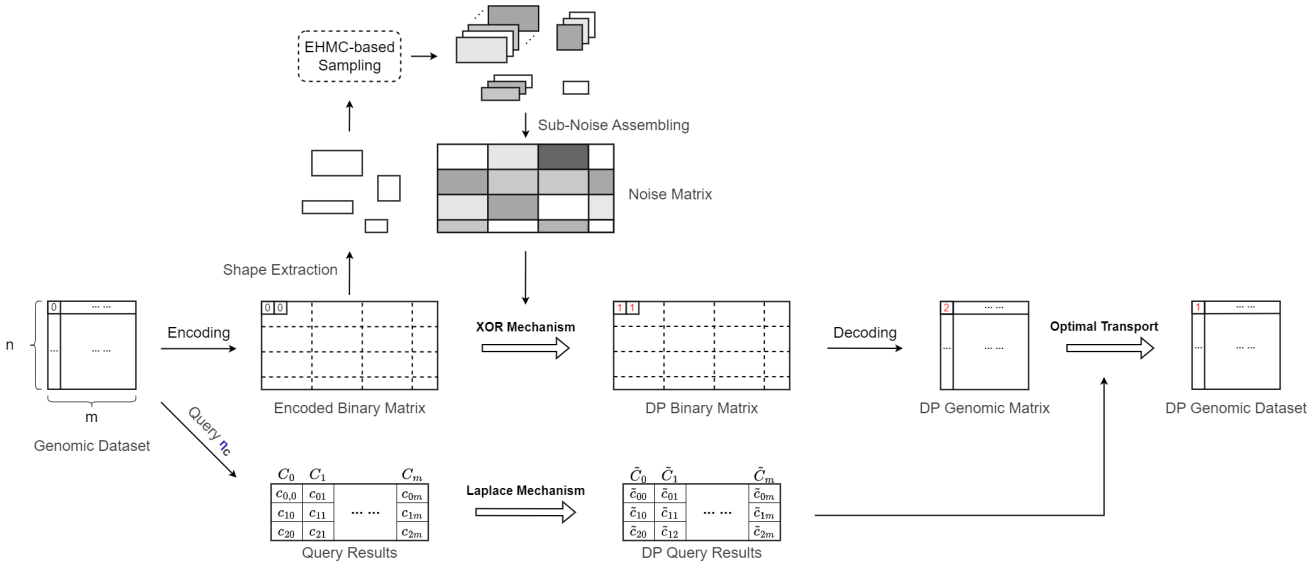


Fig. 1: The system workflow. 1) Encode the genomic dataset D to binary form D^b. 2) Divide the binary matrix into sub-matrices and record the shapes with their counts. 3) Generate noise matrices of the shapes using Exact Hamiltonian Monte Carlo (EHMC)-based sampling. 4) Construct the whole noise matrix by randomly assigning the sub-noise matrices. 5) Produce the private matrix $\hat{D}^b$ using the XOR mechanism. 6) Decode $\hat{D}^b$ to $\hat{D}$. 6) Calculate the query results C of η_c and apply Laplacian noise to each result as $\tilde{C}$. 7) Perform optimal transport on $\hat{D}$ based on $\tilde{C}$ and generate D' as the shared dataset.

considered encoding approach captures the biological property of SNP. The XOR mechanism outputs matrix- and binary-valued noise matrices by XORing the encoded dataset with the generated noise matrices. During noise generation, we calibrate the noise distribution parameter using a publicly available datasets of the same nature to preserve the inherent correlation among SNPs. Observing high time complexity of the XOR mechanism, we improve its noise generation process. In particular, after dividing the entire matrix into same-sized sub-matrices, we generate limited noise matrices and XOR each sub-matrix with a randomly chosen noise matrix among them. This optimization decreases the time cost of the XOR mechanism and becomes an essential building block for genomic data generation in practice.

The XOR mechanism is designed to protect the privacy of single binary entry in a dataset, so the privacy is guaranteed at the entry level (which means it is vulnerable to membership inference attacks, a more common attack in genomic dataset sharing). To make the XOR mechanism be robust against membership inference, it will introduce significant amount of noise. Thus, we devise a novel scheme to restore data utility. In the utility restoration stage, we first issue a count query at each SNP position in the dataset that counts the SNP value distribution and apply the Laplace mechanism [3] to ensure differential privacy. Then, for each SNP, we calculate the minimal SNP modifications needed to align the distribution with the differentially private count results. We calculate the modification plan by performing optimal transport on two distributions and modify the SNPs according to the plan.

IV. Evaluation

We implement our scheme on two realistic genomic datasets, which contain lactose-intolerant and brown-eye individuals, respectively, that are extracted from the OpenSNP [2] project. We evaluate its performance compared to two existing methods (i.e., DPSyn [6] and PrivBayes [7]) regarding GWAS reproducibility, other data utility metrics, resistance against membership inference attacks, and running time. Through the experiments, our scheme outperforms the two methods in both utility and privacy while achieving lower time complexity.

V. Conclusion

In this paper, we have proposed a novel scheme that shares genomic datasets in a privacy-preserving manner for reproducibility of genomic research outcomes. In future work, we will work on increasing reproducibility for genomic studies apart from GWAS, such as transcriptome-wide association study, genetic epidemiology, and gene–environment interaction.

References

[1] Nist 2018: Differential privacy synthetic data challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic, 2022. [Online; accessed Dec-24-2022].

[2] The opensnp project. https://opensnp.org/, 2022. [Online; accessed May-22-2022].

[3] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[4] Anisa Halimi, Leonard Dervishi, Erman Ayday, Apostolos Pyrgelis, Juan Ramón Troncoso-Pastoriza, Jean-Pierre Hubaux, Xiaoqian Jiang, and Jaideep Vaidya. Privacy-preserving and efficient verification of the outcome in genome-wide association studies. *Proceedings on Privacy Enhancing Technologies*, 2022:732–753, 07 2022.

[5] Tianxi Ji, Pan Li, Emre Yilmaz, Erman Ayday, Yanfang Ye, and Jinyuan Sun. Differentially private binary-and matrix-valued data query: an xor mechanism. *Proceedings of the VLDB Endowment*, 14(5):849–862, 2021.

[6] Ninghui Li, Zhikun Zhang, and Tianhao Wang. Dpsyn: Experiences in the nist differential privacy data synthesis challenges. *arXiv preprint arXiv:2106.12949*, 2021.

[7] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41, 2017.

Full paper: https://arxiv.org/abs/2209.06327

Poster: Reproducibility-Oriented and Privacy-Preserving Genomic Dataset Sharing

Yuzhou Jiang
yxj466@case.edu
Case Western Reserve University

Tianxi Ji
tiji@ttu.edu
Texas Tech University

Erman Ayday
exa208@case.edu
Case Western Reserve University

Motivation

- Development in genome sequencing has brought tremendous research opportunities in the genomic field.
- However, genomic research outcomes are hardly reproducible due to absence of the genomic datasets, i.e., Single Nucleotides Polymorphism (SNP) datasets.
- Existing methods under differential privacy either suffer from utility loss or high time cost, thus impractical for use.

Background

- Single Nucleotides Polymorphism (SNP) is the most common genetic variation in genetic representation.
- A SNP value represents the number of minor alleles at a position, and it can be 0, 1, or 2.
- Genome-wide association studies (GWAS) are a popular method to analyze the correlations between genetic variations and a specific trait/phenotype.

Challenges

- Sharing genomic datasets instead of sharing statistics.
- Unique utility requirement: $\chi 2$ test and odds ratio test in Genome-wide association studies (GWAS).
- Offering high data utility in other utility metrics.
- High privacy guarantee against membership inference attacks, i.e., machine learning (ML)-based attacks.
- Low computational complexity.

System and Threat Model

- Two parties: the researcher and the verifier.
  - The researcher shares the dataset that is used in their genomic research using a differentially private scheme
  - The verifier validate the research outcomes by tolerating some loss due to usage of privacy-preserving schemes.
- The attack can perform membership inference attacks:
  - Hamming distance attacks
  - Machine learning-based attacks using support vector machine and random forest.

Datasets

2 genomic datasets extracted from the OpenSNP project:
- **LACTOSE**: contains 60 lactose-intolerant individuals with 9091 SNPs
- **EYE**: contains 401 brown-eye individuals with 28396 SNPs

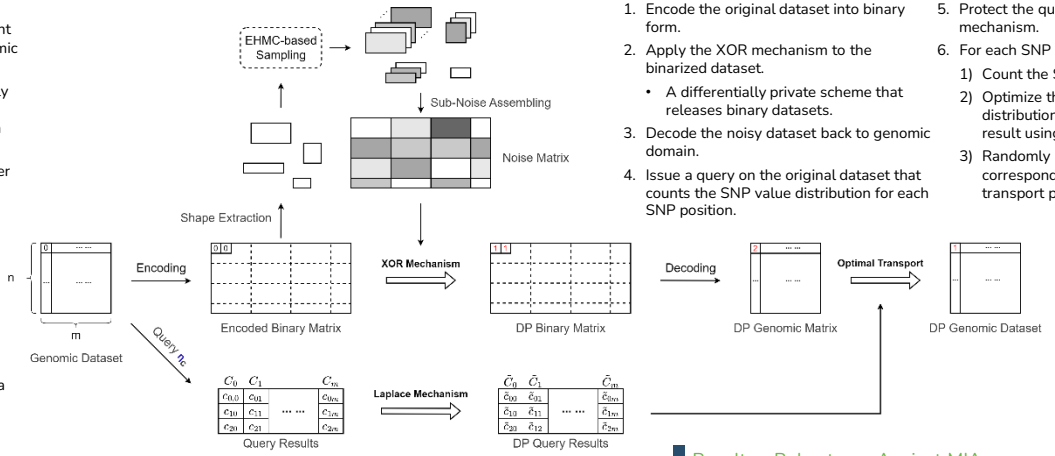
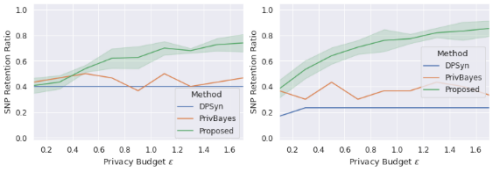


Figure 1: **System overflow**

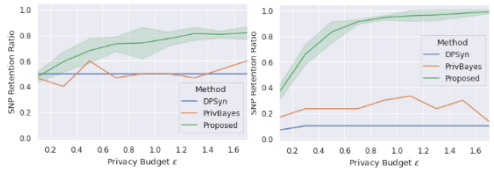
Methodology

1. Encode the original dataset into binary form.
2. Apply the XOR mechanism to the binarized dataset.
   - A differentially private scheme that releases binary datasets.
3. Decode the noisy dataset back to genomic domain.
4. Issue a query on the original dataset that counts the SNP value distribution for each SNP position.
5. Protect the query results via the Laplace mechanism.
6. For each SNP position in the noisy dataset
   1) Count the SNP distribution.
   2) Optimize the plan that adjusts distribution to match the noisy query result using optimal transport.
   3) Randomly select and flip the corresponding SNPs according to the transport plan.

Results – GWAS Reproducibility

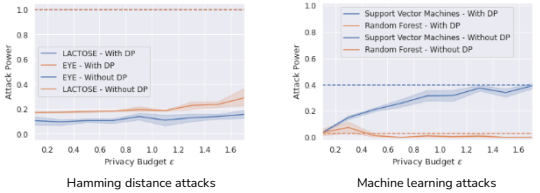


$\chi 2$ on LACTOSE | $\chi 2$ on EYE



Odds ratio on LACTOSE | Odds ratio on EYE

Results – Robustness Against MIAs



Hamming distance attacks | Machine learning attacks

Results – Running Time

Results – Other Data Utility Metrics

		$\epsilon = 0.3$			$\epsilon = 0.7$			$\epsilon = 1.1$			$\epsilon = 1.5$		
		Proposed	DPSyn	PrivBayes	Proposed	DPSyn	PrivBayes	Proposed	DPSyn	PrivBayes	Proposed	DPSyn	PrivBayes
LACTOSE	Point Error	**0.95**	1.29	1.30	**0.85**	1.29	1.28	**0.77**	1.29	1.29	**0.68**	1.29	1.26
	Sample Error	**0.56**	0.84	0.85	**0.48**	0.84	0.84	**0.43**	0.84	0.84	**0.37**	0.84	0.83
	Mean Error	**0.06**	0.50	0.51	**0.03**	0.50	0.48	**0.01**	0.50	0.49	**0.01**	0.50	0.47
	Variance Error	**0.02**	0.26	0.19	**0.02**	0.26	0.18	**0.01**	0.26	0.19	**0.01**	0.26	0.20
EYE	Point Error	**0.51**	1.07	0.97	**0.47**	1.07	1.03	**0.43**	1.07	0.99	**0.40**	1.07	0.94
	Sample Error	**0.30**	0.71	0.64	**0.26**	0.71	0.67	**0.24**	0.71	0.67	**0.22**	0.71	0.61
	Mean Error	**0.01**	0.54	0.46	**0.00**	0.54	0.49	**0.00**	0.54	0.49	**0.00**	0.54	0.41
	Variance Error	**0.02**	0.31	0.23	**0.01**	0.31	0.21	**0.01**	0.31	0.24	**0.00**	0.31	0.24

Future Work

- Increase reproducibility for genomic studies apart from GWAS, e.g.,
  - transcriptome-wide association study
  - genetic epidemiology
  - gene–environment interaction
- Explore the feasibility of applying other schemes for better performance.
- Achieve both liability and privacy guarantees during genomic dataset sharing by incorporating dataset fingerprinting techniques.

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.