Poster: Local Differentially-Private Genomic Database Fingerprinting

Tianxi Ji
Texas Tech
University
tiji@ttu.edu

Erman Ayday
Case Western
Reserve University
exa208@case.edu

Emre Yilmaz
University of
Houston-Downtown
yilmaze@uhd.edu

Pan Li
Case Western
Reserve University
lipan@case.edu

***Abstract*—Sharing genomic databases is critical to the collaborative research in computational biology. A shared database is more informative than specific GWAS statistics as it enables "do-it-yourself" calculations. Constructed genomic databases involve intellectual efforts from the curator and sensitive information of participants, thus when sharing the databases with others, the curator (database owner) should be able to prevent unauthorized redistributions and protect individuals' genomic data. As it becomes increasingly common for a same database be shared with multiple recipients, the shared database should also be robust against collusion attack, where malicious recipients combine their individual versions of received database to forge a pirated copy with the hope that none of them can be traced back.**

We propose to develop a novel steganographic scheme to achieve both copyright protection and privacy preservation guarantees for the shared genomic databases by inserting random fingerprint bit-strings (steganographic marks) into each shared copy. In a nutshell, the proposed scheme attains local differential privacy of fingerprinted genomic databases by leveraging the intrinsic randomness of the fingerprint marks. To defend against collusion attacks, we augment Gen-Scope with the Tardos codes, which is a powerful traitor tracing techniques to mitigate collusion attacks by identifying the colluders with high probability.

I. Introduction

In recent decades, gigantic amounts of genomic data have been generated and collected at a unprecedented rate in Genome-wide association studies (GWAS) to discover the associations between phenotypes and particular traits or human diseases. Moreover, the implementation and sharing of genomic databases, e.g., the single nucleotide polymorphism database (dbSNP) [1], has significantly advanced the collaborative research on physical mapping, population genetics, human biology, and modern medicine.

While the benefits of collecting genomes and constructing genomic databases are trumpeted by the computational biology community, the increased availability of such data has raised concerns about the data owners' copyright and the data contributors' privacy. Thus, an owner of genomic database will only share its data to authorized recipients, e.g., service providers (SPs) like hospitals and research institutions to prevent illegal redistribution of data. On the other hand, genomic data contains sensitive features that can be used to identify individuals (via forensics), connect to other family members (via kinship), and infer individuals' health condition (associating SNPs with diseases) [2]. The database owner is also obligated to protect the privacy of the individuals (data contributors). Besides, it is noteworthy that the GDPR lists genetic data as "special categories of personal data" that is subject organizational and technical safeguards.

The motivation of this work is to develop a feasible genomic database sharing scheme to provide researchers access to genomic data for the purposes of collaborative research and "do-it-yourself" calculations. The developed scheme should have both copyright protection and privacy preservation guarantees for the shared genomic databases. We focus on the sharing of the entire genomic database instead of answering specific GWAS statistics (e.g., correlation between genome pairs). This is because in a typical GWAS process, the researchers do not know in advance which genome pairs to use and what types of statistical tests to query [3].

II. System, Properties, Threats, and Methods

In general, a genomic database recipient (SP) can be any of the following: (1) an honest party who will use the received dbSNP to perform GWAS, (2) an attacker who wants to make illegal profits by changing some entries in its received database and making pirate copies of it, or (3) a curious party who tries to infer the original genome values. Thus, our system is designed to achieve the following properties

- (i) high utility for the fingerprinted database in order to support accurate GWAS,
- (ii) copyright protection to discourage illegal redistribution, i.e., to successfully extract a malicious SP's fingerprint (even if it tries to distort the fingerprint in its received database to mitigate detection), when identifying a pirated version of the released database,
- (iii) local differential privacy guarantee against attributes inference attacks, i.e., a data analyst cannot infer the original value of each genomic data.

From the perspective of malicious SPs, their goals are to

- (a) illegally redistribute received genomic database (i.e., make pirated copies by launching various attacks targeting the inserted fingerprint bit-string) without being accused by the database owner, **and/or** launch inference attack aiming to recover the original values of genomc in its received copy,
- (b) preserve database utility to gain illegal profit.

We show the overview of the system model (using dbSNP as an example) in Figure 1. We represent the genomic database owned by Alice as $\mathbf{R}$. She wants to share $\mathbf{R}$ with N SPs (e.g., to receive specific services, to help researchers, or for collaborative research). To prevent unauthorized redistribution of the database by a malicious SP, Alice embeds unique fingerprints in all shared copies of the dbSNP. The fingerprint essentially changes different entries in $\mathbf{R}$ at different SNP positions (indicated by the yellow dots). The fingerprint (a binary bit-string) generated for the ith SP (SP_i) is denoted as f_{SP_i}, and the dbSNP received by SP_i is represented as $\widetilde{\mathbf{R}}_{\mathrm{SP_i}}$. In Figure 1, if SP_i forges a pirated dbSNP, i.e., $\overline{\mathbf{R}}$, by changing some values (indicated by the red dots) in its received copy, i.e., $\widetilde{\mathbf{R}}_{\mathrm{SP_i}}$, Alice will be able to accuse SP_i for data leakage with high probability by extracting f_{SP_i} from $\overline{\mathbf{R}}$. In addition to the copyright protection, Alice also preserves the privacy of SNP data and maintain high utility for the shared database.

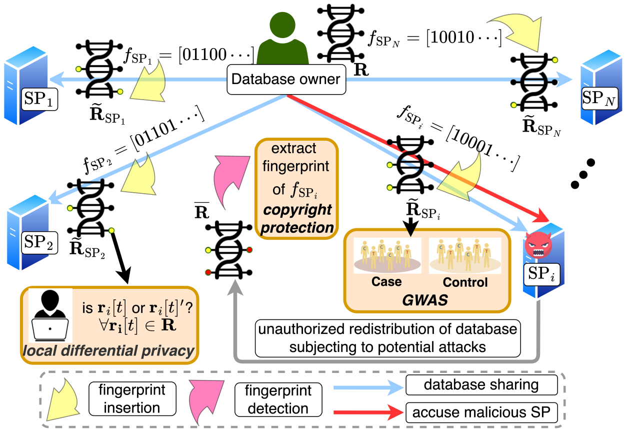

Database owner
$\mathbf{R}$
$f_{\mathrm{SP}_1} = [01100\cdots]$
$f_{\mathrm{SP}_N} = [10010\cdots]$
SP$_1$
SP$_N$
$\widetilde{\mathbf{R}}_{\mathrm{SP}_1}$
$\widetilde{\mathbf{R}}_{\mathrm{SP}_N}$
$f_{\mathrm{SP}_2} = [01101\cdots]$
$f_{\mathrm{SP}_i} = [10001\cdots]$
extract fingerprint of f_{SP_i} copyright protection
$\overline{\mathbf{R}}$
$\widetilde{\mathbf{R}}_{\mathrm{SP}_i}$
SP$_2$
$\widetilde{\mathbf{R}}_{\mathrm{SP}_2}$
Case
Control
GWAS
SP$_i$
is $\mathbf{r}_i[t]$ or $\mathbf{r}_i[t]'$? $\forall \mathbf{r_i}[t] \in \mathbf{R}$
local differential privacy
unauthorized redistribution of database subjecting to potential attacks
fingerprint insertion
fingerprint detection
database sharing
accuse malicious SP

Fig. 1: Alice adds a unique fingerprint in each copy of her dbSNP $\mathbf{R}$ when sharing. The inserted fingerprint changes entries at different locations (the yellow dots) in $\mathbf{R}$. She is able to identify the malicious SP who pirates and redistributes her database using a distorted fingerprint. All shared dbSNP copies achieve local DP and fingerprint robustness.

Methods. To achieve both copyright protection and privacy preservation for genomic databases, a straightforward two-step approach is to insert fingerprint into a differential-privately sanitized genome. However, it will significantly reduce the utility of the final genomic databases, because it requires adding separate noises to achieve the two guarantees; first adding noise to attain privacy guarantee, then, adding additional noise (via fingerprinting) to obtain copyright guarantee.

By building upon our previous solutions [4], [5], [6], we propose **Gen-Scope**, which shares **gen**omic databases and **s**imultaneously achieve **co**pyright protection and **p**rivacy pres**e**rvation via one-shot noise (fingerprint) insertion. The fingerprint is a binary and steganographic marks, and is obtained by a message authentication code (MAC) involving a cryptographic hash function, a secret cryptographic key and the identity of the recipient. In Gen-Scope, the inserted fingerprint can also be used to protect the privacy of the genomic data. The key idea is to leverage the intrinsic randomness of fingerprint insertion and transform it into a provable privacy guarantee. In particular, we first observe that fingerprint insertion essentially flips each bit of a genomic data randomly, and this leads to the value of that genome being changed with certain probability, then from which, we can derive the privacy guarantee in the form of local differential privacy (LDP) [7]. Since Gen-Scope only inserts noise once, the final database will have high utility.

As it has been increasingly common for a database owner to share its data with various recipients. The inserted fingerprints may still be compromised by collusion attack, where multiple malicious recipients combine their individual versions of fingerprinted databases to forge a pirated copy with the hope that none of them can be traced back. To achieve robustness against collusion attacks, we improve Gen-Scope by incorporating the Tardos code [8], which is one of the most powerful techniques to fight against collusion attacks by identifying the colluders with very high probability.

Additionally, we will also analyze the required fraction of changed genomic data entries for our proposed Gen-Scope and the two-stage approach (local differentially private perturbation followed by fingerprinting) to achieve the required privacy and copyright guarantees. Then, we will theortically establish the relationships between privacy, utility, and fingerprint robustness, i.e., derive the closed-from expression between ϵ (privacy budget), robustness metrics, database utility metrics, and the number of colluders participating in collusion along with the probabilities of falsely accuse innocent SPs and failed to accuse one of the colluders.

III. Conclusion

In this poster, we have proposed Gen-Scope, which is the first genomic database fingerprinting scheme that can simultaneously achieve copyright protection, privacy preservation, and accurate value (utility) when sharing genomic databases. Gen-Scope attains LDP by leveraging the intrinsic randomness during fingerprint insertion. We also discussed how to improve Gen-Scope to defend against collusion attacks.

References

[1] Stephen T Sherry, Minghong Ward, and Karl Sirotkin. dbsnp—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research*, 9(8):677–679, 1999.

[2] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):1–44, 2015.

[3] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[4] Tianxi Ji, Erman Ayday, Emre Yilmaz, and Pan Li. Robust fingerprinting of genomic databases. In *30th International Conference on Intelligent Systems for Molecular Biology*, ISMB'21, Oxford, England, 2021. Oxford University Press.

[5] Tianxi Ji, Emre Yilmaz, Erman Ayday, and Pan Li. The curse of correlations for robust fingerprinting of relational databases. In *24th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID '21, page 412–427, 2021.

[6] Tianxi Ji, Erman Ayday, Emre Yilmaz, and Pan Li. Towards robust fingerprinting of relational databases by mitigating correlation attacks. *IEEE Transactions on Dependable and Secure Computing*, 2022.

[7] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[8] G Tardos. Optimal probabilistic fingerprint codes. In *Proc. of the 35th annual ACM symposium on theory of computing, San Diego, CA, USA*, pages 116–125, 2005.

Local Differentially-Private Genomic Database Fingerprinting

Tianxi Ji[1], Erman Ayday[2], Emre Yilmaz[3], and Pan Li[2]

[1]Texas Tech University, Lubbock, TX, USA, [2]Case Western Reserve University, Cleveland, OH, USA, and [3]University of Houston-Downtown, Houston, TX, USA.

Motivation

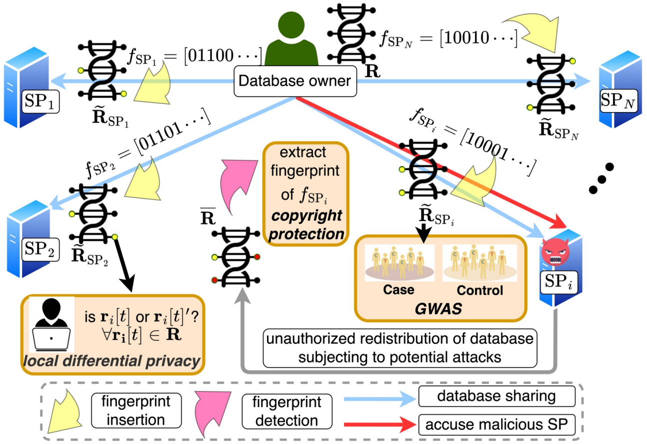
- Sharing genomic databases is critical to the collaborative research in computational biology. Shared databases are more informative than GWAS statistics as it enables “do-it-yourself" calculations.

- Constructed genomic databases involve intellectual efforts from the curator and sensitive information of participants, thus when sharing the databases with others, the curator (database owner) should be able to
  1. maintain high utility for the shared database in order to support accurate GWAS,
  2. protect copyright to discourage illegal redistribution, trace the source of data leakage, and accuse malicious party with evidence
  3. preserve the privacy of genomic data entries.

- A malicious genomic database receiver may want to
  1. illegally redistribute received copy without being accused by the owner, and/or launch inference attack aiming to recover the original values of genome,
  2. preserve database utility to gain illegal profit.

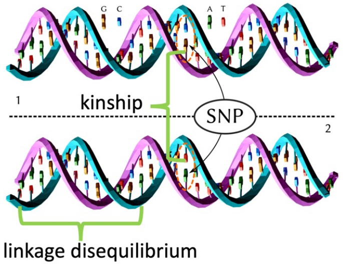
- Develop a novel **privacy-preserving steganography** scheme which achieves both privacy and liability guarantees on the shared genomic database while only **slightly** reduces the genomic database utility.

Proposed System Model


$f_{\mathrm{SP}_1} = [01100\cdots]$
Database owner
$\mathbf{R}$
$f_{\mathrm{SP}_N} = [10010\cdots]$
SP_1
$\widetilde{\mathbf{R}}_{\mathrm{SP}_1}$
SP_N
$\widetilde{\mathbf{R}}_{\mathrm{SP}_N}$
$f_{\mathrm{SP}_2} = [01101\cdots]$
$f_{\mathrm{SP}_i} = [10001\cdots]$
extract fingerprint of f_{SP_i}
copyright protection
$\overline{\mathbf{R}}$
$\widetilde{\mathbf{R}}_{\mathrm{SP}_i}$
SP_2
$\widetilde{\mathbf{R}}_{\mathrm{SP}_2}$
Case
Control
GWAS
SP_i
is $\mathbf{r}_i[t]$ or $\mathbf{r}_i[t]'$? $\forall \mathbf{r_i}[t] \in \mathbf{R}$
local differential privacy
unauthorized redistribution of database subjecting to potential attacks
fingerprint insertion
fingerprint detection
database sharing
accuse malicious SP

Threats against steganographic marks

- Simple attacks: random bit flipping, subset (superset) attack
- Correlation attack [1,2]: leveraging the correlation among genomic data entries



- Collusion attack [3]: malicious recipients combine their individual versions of received database to forge a pirated copy with the hope that none of them can be traced back

Methodology

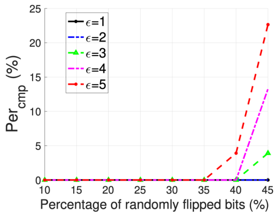
- Leverage the intrinsic randomness of fingerprint insertion and transform it into a provable privacy guarantee.
  - fingerprint insertion essentially flips each bit of a SNP data randomly, and this leads to the value of that SNP being changed with certain probability
  - connect randomness to LDP

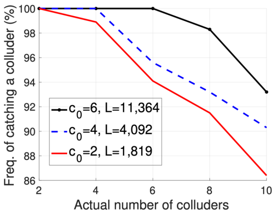
- Incorporate the Tardos code [4], one of the most powerful techniques to fight against collusion attacks by identifying the colluders with very high probability

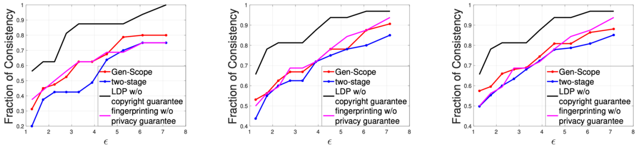
```
1 Sample a random variable $p$ from probability density function
      $f(p|t) = \frac{1}{2\arcsin(1-2t)} \frac{1}{\sqrt{p(1-p)}}, t \in (0, 0.5).$
2 Generate the Tardos fingerprint string, i.e., $\mathbf{f} \sim \mathit{Bernoulli}(p)$.
```

- Calibrate t with ϵ (LDP budget)
- Connect database utility, t, ϵ, robustness metrics, colluder size, probability of innocent party being falsely accused, and probability of one colluder is not accused

Preliminary results







(a) top 10% SNPs (b) top 20% SNPs (c) top 30% SNPs

Conclusion and Future Work

- Propose the first privacy-preserving genomic database fingerprinting scheme
- Robust against collusion attack
- Outperform two-stage approaches

- Cumulative privacy loss control when database is repeatedly shared
- Utility boosting considering biological features

References

[1] Tianxi Ji, Emre Yilmaz, Erman Ayday, and Pan Li. The curse of correlations for robust fingerprinting of relational databases. In 24th International Symposium on Research in Attacks, Intrusions and Defenses, RAID ’21, page 412–427, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Tianxi Ji, Erman Ayday, Emre Yilmaz, and Pan Li. Robust fingerprinting of genomic databases. In 30th International Conference on Intelligent Systems for Molecular Biology, ISMB’21, Oxford, England, 2021. Oxford University Press.

[3] Yingjiu Li, Vipin Swarup, and Sushil Jajodia. Fingerprinting relational databases: Schemes and specialties. IEEE Transactions on Dependable and Secure Computing, 2(1):34–45, 2005.

[4] G Tardos. Optimal probabilistic fingerprint codes. In Proc. of the 35th annual ACM symposium on theory of computing, San Diego, CA, USA,pages 116–125, 2005.

Contact: Tianxi Ji, CS department, Texas Tech University, tiji@ttu.edu