DyPS: Dynamic, Private and Secure GWAS

Talk proposal - GenoPri 2021

Túlio Pascoal[*], Jérémie Decouchant[†], Antoine Boutet[‡], Paulo Esteves-Verissimo[§]
[*]University of Luxembourg, [†]Delft University of Technology, [‡]Univ. Lyon, INSA Lyon, Inria, CITI, [§]KAUST - RC3

***Abstract*—Genome-Wide Association Studies (GWAS) identify the genomic variations that are statistically associated with a particular phenotype (e.g., a disease). GWAS results, i.e., statistics, benefit research and personalized medicine. The confidence in GWAS increases with the number of genomes analyzed, which encourages federated computations where biocenters periodically include newly sequenced genomes. However, for legal and economical reasons, this collaboration can only happen if a release of GWAS results never jeopardizes the genomic privacy of data donors, if biocenters retain ownership and cannot learn each others' data. Furthermore, given the reduced cost of sequencing DNA nowadays, there is now a need to update GWAS results in a dynamic manner, while enabling donors to withdraw consent at any time. Therefore, two challenges need to be simultaneously addressed to enable federated and dynamic GWAS: (i) the computation of GWAS statistics must rely on secure and privacy-preserving methods; and (ii) GWAS results that are publicly released should not allow any form of privacy attack. In this talk, we will introduce the problem we consider in more details and present DyPS [1], the framework we have designed and recently presented at the Privacy Enhancing Technologies Symposium (PETS). We refer the reader to the full paper[1] for the details we cannot cover in this short version.**

I. Introduction

GWAS allow the discovery of correlations between genetic variations, such as single-nucleotide polymorphisms (SNPs), and a phenotypic trait. For example, GWAS are widely used for the early detection of disease susceptibility. GWAS produce two types of statistics to identify correlations between SNPs and phenotypes: (i) *aggregate* statistics, such as minor allele frequencies (MAF), single and pairwise allele frequencies; and (ii) *test* statistics, such as linkage disequilibrium (LD), χ^2 and *p-values*.

To obtain meaningful statistical findings, GWAS require large datasets, which motivates the development of federated environments where several biocenters can securely and collaboratively compute GWAS statistics over their genomic data. Nevertheless, enforcing secure distributed GWAS computation alone is not enough since the publication of GWAS results might enable sophisticated membership [2]–[4] or recovery attacks [5]–[7]. Recent works have focused on either secure and privacy-preserving GWAS computation, or on privacy-preserving GWAS data releases, but not on both. However, we argue that those two problems need to be simultaneously addressed, since both can produce information leakage and lead to privacy attacks.

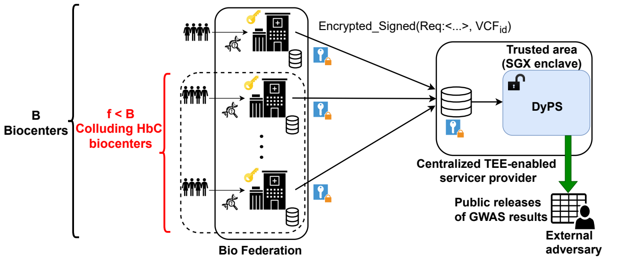

Fig. 1: DyPS' architecture to enable secure and privacy-preserving federated GWAS.

In addition, we observe that attacks can be facilitated by collusions among dataholders. Indeed, a subset of biocenters might be able to collude to infer other biocenters' genomic data. Furthermore, there is now a need to enable dynamic GWAS where new genomes can be added or removed (in order to comply with privacy regulations, such as GDPR) over time. If not crafted with proper care, GWAS updates might be combined by adversaries and become subject to privacy attacks.

We introduce DyPS [1], a novel framework that provides both secure GWAS computation and privacy-preserving public releasing of GWAS results. DyPS leverages a Trusted Execution Environment to secure dynamic GWAS computations. Moreover, DyPS uses a scaling mechanism to speed up the releases of GWAS results according to the evolving number of genomes used in the study, even if individuals retract their participation consent. Lastly, DyPS also tolerates up to all-but-one biocenters colluding without privacy leaks.

II. Privacy-preserving GWAS computation

DyPS leverages a TEE-enabled centralized server to allow the safe sharing of genomic data from several biocenters, and therefore allows individuals to stay in control of their own genomes, which are never revealed to other biocenters. DyPS's computations are performed inside an enclave, which is an processing environment that cannot be tampered with. Even privileged code (e.g., the hosting operating system) cannot access its data. DyPS uses an Intel SGX enclave that can be attested by biocenters to certify that trusted code is being run by a genuine SGX enclave. Once the remote attestation step succeeds, biocenters can securely outsource their genome data through authenticated secure channels, after encrypting and signing it. Moreover, it is also possible to store data outside the enclave (which only has 96 MB of available memory) by using a sealing mechanism. Such data is encrypted by the enclave and can only be accessed and verified by this enclave.

We illustrate DyPS's federated architecture in Figure 1. We assume the presence of B honest-but-curious biocenters,

[1] https://petsymposium.org/2021/files/papers/issue2/popets-2021-0025.pdf

among which $f = B - 1$ might collude. DyPS's algorithms only run inside the enclave, which enforces isolation, confidentiality, and integrity of both the data and operations. The enclave periodically collects the requests from the various biocenters and decides which requests are to be executed to dynamically update and release the GWAS results.

III. Protecting dynamic releases of GWAS

DyPS mainly builds on Zhou et al.'s work [5] and on SecureGenome [8], which define conditions for identifying when the release of static GWAS results is safe. In particular, DyPS extends such approaches in order to support dynamic updates of GWAS results.

To protect GWAS aggregate statistics against recovery attacks and test statistics against recovery and membership attacks, DyPS makes sure that enough genomes are being used so that any probabilistic polynomial-time (p.p.t.) adversary observing released results cannot successfully recover the complete solution space, and determine the right genotype sequences within the human-genome data set (i.e., individuals' real genome sequence) that participated in a given release.

In addition, to protect aggregate statistics against membership attacks, DyPS uses the Likelihood Ratio Test (LR-Test) to certify that no genomes participating in a release can be identified. For that purpose, DyPS leverages SecureGenome [8], which is an LR-test approach that defines over which SNP positions can allele frequencies be safely released within a cohort of genomes.

In order to allow safe dynamic releases of GWAS, DyPS adopts the following pipeline:

(1) FIFO request pool. DyPS collects and treats genome addition and removal requests from biocenters in FIFO order.

(2) Requests selection. DyPS selects a safe batch of genome requests so that any possible subset of honest biocenters collectively have enough operations. This way, an external adversary or the remaining possibly colluding biocenters cannot isolate and attack a small number of requests.

(3) GWAS processing. DyPS computes and releases test statistics over the selected genomes.

(4) Membership tests. DyPS identifies over which subset of SNPs can aggregate statistics also be released or updated.

IV. Performance evaluation

We used Graphene to run DyPS inside an SGX enclave. Biocenters adopted AES256 encryption and ECDSA signatures to outsource genome data. Genome addition or removal requests are generated following a Poisson distribution, with $\lambda = 8$ for additions and $\lambda = 6$ for removals. For larger-scale experiments, we proportionally increased λ. We evaluated DyPS under different GWAS scenarios. For instance, we considered a number of biocenters that ranges from 4 to 7, with up to $f = B - 1$ of them colluding. We used simulated genomes (up to 6 million) and real genomes (up to 27,895 genomes).

First, we show that DyPS request selection mechanism runs with O(1) complexity while a Brute Force approach would run in O(2^N). In addition, DyPS requires $\approx 75KB$ to encode 300,000 SNPs, and the overall request size is 258 bits, i.e., ≈ 48 Bytes after encryption. DyPS also presents a reasonable memory consumption (2 MB on average, which is within SGX's limits). DyPS also remains practical under a large-scale setting (5 biocenters, 5,000 SNPs, and approximately 28,000 real genomes), with a reasonable running time, even though it would increase with the number of releases due to the membership evaluation.

DyPS maintains genomic privacy, but releases GWAS results less frequently than a naive approach, which in turn would leave 4.98% of its releases vulnerable to attacks and 8% of the genomes exposed to privacy threats. In addition, the existence of colluding players only impacts the speed with which requests are selected and processed. As expected, the larger the number of colluding players the larger are the addition and removal request delays. Last but not least, we show that DyPS can regularly release GWAS results thanks to our dynamic scaling methods.

V. Conclusion

This paper aims at reconciling secure GWAS processing with privacy-preserving GWAS releases. On one hand, an adversary might launch privacy attacks during the computation of GWAS. On the other hand, GWAS releases must preserve privacy, which requires computations over the distributed genomic dataset. Unfortunately, current works in the literature are not able to combine both aspects of a fully privacy-preserving GWAS (processing and releasing) in a homogenized form. Moreover, due to cheaper costs for DNA sequencing and to comply with data-privacy regulations, where donors should be able to withdraw consent at any time, GWAS results need to be safely updated on the fly. In this regard, we presented DyPS, a TEE-based framework that enables privacy-preserving processing and dynamic releasing of federated GWAS. DyPS is the first work capable of reconciling both challenges in a secure, scalable, and reliable fashion, and thus of protecting genomic privacy in terms of both biocenters' and donors' genome data. Future works include considering the existence of byzantine biocenters, and an oblivious version of DyPS to cope with side-channel attacks against SGX enclaves.

References

[1] T. Pascoal, J. Decouchant, A. Boutet *et al.*, "Dyps: Dynamic, private and secure gwas," *PoPETS*, 2021.

[2] N. Homer, S. Szelinger, M. Redman *et al.*, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS genetics*, 2008.

[3] K. B. Jacobs, M. Yeager, S. Wacholder *et al.*, "A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies," *Nature genetics*, vol. 41, no. 11, p. 1253, 2009.

[4] M. Humbert, K. Huguenin, J. Hugonot *et al.*, "De-anonymizing genomic databases using phenotypic traits," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 2, 2015.

[5] X. Zhou, B. Peng, Y. F. Li *et al.*, "To release or not to release: Evaluating information leaks in aggregate human-genome data," in *Esorics*, 2011.

[6] M. Humbert, E. Ayday, J.-P. Hubaux *et al.*, "Quantifying interdependent risks in genomic privacy," *ACM Transactions on Privacy and Security (TOPS)*, vol. 20, no. 1, pp. 1–31, 2017.

[7] E. Ayday and M. Humbert, "Inference attacks against kin genomic privacy," *IEEE Security & Privacy*, vol. 15, no. 5, pp. 29–37, 2017.

[8] S. Sankararaman, G. Obozinski, M. I. Jordan *et al.*, "Genomic privacy and limits of individual detection in a pool," *Nature genetics*, vol. 41, no. 9, pp. 965–967, 2009.