Towards practical Genome-Wide Association Studies: Overview and Challenges

Túlio Pascoal
University of Luxembourg
tulio.pascoal@uni.lu

Jérémie Decouchant
Delft University of Technology
j.decouchant@tudelft.nl

Marcus Völp
University of Luxembourg
marcus.voelp@uni.lu

ABSTRACT

The popularization of large-scale federated Genome-Wide Association Study (GWAS) where multiple data owners share their genome data to conduct federated analytics uncovers new privacy issues that have remained unnoticed or not given proper attention. Indeed, as soon as a diverse type of interested parties (e.g., private or public biocenters and governmental institutions from around the globe) and individuals from heterogeneous populations are participating in cooperative studies, interdependent and multi-party privacy appear as crucial issues that are currently not adequately assessed. In fact, in federated GWAS environments, the privacy of individuals and parties does not depend solely on their own behavior anymore but also on others, because a collaborative environment opens new credible adversary models. For instance, one might want to tailor the privacy guarantees to withstand the presence of potentially colluding federation members aiming to violate other members' data privacy and the privacy deterioration that might occur in the presence of interdependent genomic data (e.g., due to the presence of relatives in studies or the perpetuation of previous genomic privacy leaks in future studies). In this work, we catalog and discuss the features, unsolved problems, and challenges to tackle toward truly end-to-end private and practical federated GWAS.

1. INTRODUCTION

In this work, we identify and discuss current and future challenges to pave the way for end-to-end private and accessible to the masses Genome-Wide Association Study (GWAS), which we designate as practical GWAS. In particular, we argue that most of the challenges are consequences of a trend towards interdependent and multi-party genomic settings.

In the following, we introduce the state of the art of GWAS, discussing the benefits, drawbacks and remaining challenges of existing mechanisms that allow privacy-preserving GWAS.

GWAS is an observational study that computes correlation statistics to identify associations between genomic variations with a particular phenotype, e.g., a disease. In fact, GWAS has been playing an essential role in developing personalized and preventive medicine, early disease diagnoses, etc. GWAS operates over genome variants (usually Single-Nucleotide Polymorphisms – SNPs) of donors from a case population consisting of individuals expressing a given trait and a control population comprising healthy individuals. To improve the statistical power, and therefore the confidence of GWAS findings, the research community adopts collaborative environments where GWAS are conducted over a larger set of genomes that several stakeholders (commonly geographically distant from each other) disseminate for federated analytics. Such a scheme is known as federated GWAS [12, 23, 25].

Despite its benefits, federated GWAS raises several privacy constraints since managing, outsourcing, and processing such sensitive data needs appropriate care [9]. To that extent, many works have proposed privacy-preserving solutions that leverage cryptographic mechanisms to protect the integrity, confidentiality, and privacy of the data communicated by federation members. For instance, solutions that relies on local Differential Privacy (DP) [8], Homomorphic Encryption (HE) [12], Secure Multi-party Computation (SMC) [28] and Trusted Execution Environment (TEE) [23, 20] have been offered.

Unfortunately, relying on privacy-preserving mechanisms to allow secure and private federated GWAS *processing* is not sufficient since to achieve its full potential the results of a GWAS should ideally be published publicly, which requires additional care. In particular, public access to GWAS results has been prohibited since several works have demonstrated the feasibility of genomic privacy attacks on the observation of GWAS statistics. For instance, Zhou et al. [32] described genomic recovery attacks that can infer SNP information of participants in a study by the observation of released statistics and GWAS metadata (e.g., number of participants and SNPs). Later, other works showed how high-order correlations among variants of the human genome [29, 10] and kinship [15, 3, 16] could be leveraged to increase the power of attribute inference attacks. Similarly, Homer et al. introduced membership inference attacks that can identify the participation of a given individual in a study using likelihood-ratio statistical tests [14]. Following this rationale, subsequent works have offered several other inference methods to measure the risks of membership inference of individuals from GWAS statistics [26, 30, 32]. Coupled with the above issues, Humbert et al. not only demonstrate potential risks associated with interdependent privacy but also measure how genomic privacy is jeopardized when adversaries combine the observation of publicly known GWAS statistics with familial relationships of individuals [17, 15].

Notably, these attacks might reduce the number of donors willing to participate in future studies since they cannot trust existing solutions to ensure their privacy [11, 23]. Besides that, federation members also need assurance that their private shares cannot be attacked during the federated analysis, which consists of aggregation operations that might leak secret data. Therefore, we claim that federated

GWAS solutions must place maximum efforts on reconciling privacy-preserving *processing* and *releasing* [23], thus, implanting a holistic solution that satisfies a wider range of privacy constraints.

Although significant efforts have been placed to create end-to-end private federated GWAS workflows, new unsolved challenges arise when considering practical GWAS properties, which aspire to support current data privacy regulations requirements. Moreover, additional privacy issues arise when interdependent and multi-party privacy are considered, which we discuss next.

2. TOWARDS PRACTICAL GWAS

For the upcoming years, we envision the development of a practical GWAS environment capable of accommodating the best practices of secure and privacy-preserving *processing* with *releasing* of GWAS results, but at the same time complying with 21st-century data privacy regulations. Based on these assumptions, we categorize below requirements needed to support practical GWAS in the future.

1. Allowing public releases of results so that the benefits of GWAS findings can achieve their full potential, i.e., being delivered and accessible to the masses. We note that since the description of the first membership inference attack by Homer et al. [14], the NIH has restricted access to GWAS results [31].

2. Enabling donors to withdraw consent at any time to comply with privacy regulations, such as GDPR. However, such operations need to be carefully crafted since an improper update might facilitate genomic privacy attacks mounted on the observation of how statistics have evolved [23].

3. Considering that new donors are sequenced over time (and currently at an increased pace due to reduced DNA sequencing costs [13]), and therefore GWAS results should be updated as soon as possible, but in a private and safe fashion because result updates inherit the same issues discussed above.

4. Acknowledging that the presence of multiple studies increases the chance of overlapping data being used [22, 18], which directly impacts the privacy guarantees of existing mechanisms that aim at enforcing the privacy of single studies and therefore cannot support privacy under the presence of interdependent GWASes.

5. Assuming stronger adversary model assumptions. For instance, the presence of honest-but-curious parties that might collude with other parties in order to strengthen their knowledge to facilitate genomic privacy attacks and/or circumvent known private release conditions.

2.1 Challenges for enforcing practical GWAS

Although a considerable number of works that offer secure and privacy-preserving federated GWAS have been proposed, only a minority combines privacy-preserving processing with private GWAS releases [24, 4, 23]. Indeed, simultaneously addressing all these constraints is not an intuitive task. Further, when bearing practical GWAS features in mind, additional care needs to be enforced, which we describe below.

Besides identifying and creating the conditions to allow safe updates of GWAS results (independently of the assumed approach, e.g., relying on Differential Privacy mechanisms [24, 4], theoretical complexity analyses of attacks [32, 29], or statistical inference methods [26, 23]), dynamic releases are highly dependent on the genomic privacy deterioration over time [5]. In particular because genomic privacy degrades according to the sharing rate of genomic data and the heterogeneity of populations. Therefore, ensuring safe releases considering static privacy (i.e., examining risks until the moment results are published) might not be sufficient. As a result, additional diligence to assess the impact of a given release on subsequent ones should be investigated.

In addition, there is a known issue regarding privacy conflicts that arise from the presence of dependent records. In fact, Almadhoun et al. [1] have shown that the privacy guarantees enabled by Differential Privacy mechanisms cannot be kept when dependent genomes are present in a study. Besides that, unfortunately, we are not aware of any DP-based solution that can enforce the same DP guarantees over continuous releases of data. In contrast, genome-oriented statistical inference methods such as the one proposed by Sankararam et al. [26] enables the detection of relatives in studies, even though extensions must be compelled to preserve the privacy guarantees of the scheme under a dynamic release setting [23].

On the system side specter, collusion-tolerance remains a feature that has not been enforced by all cryptographic approaches, mainly when resulting data needs to be published. Secret sharing, threshold-SMC, or collective HE methods allow private shares to be securely and privately aggregated, e.g., not allowing improper recovery if a sufficient number of shares is gathered. These approaches assume that aggregated data is kept within a trusted curator or is secured by relying on cryptographic primitives (i.e., stored in an encrypted form and decrypted only by authorized parties). However, when results (e.g., GWAS statistics) from aggregated data needs to be disclosed (to allow open-access GWAS), they might become subject to genomic privacy attacks from external adversaries or internal (colluding) parties. On the one hand, external adversaries can mount standard recovery or membership attacks on the observation of the GWAS metadata and released statistics [32, 14]. On the other hand, colluding members of a federation might join their data in order to decrease the solution space one has to attack from the GWAS statistics when launching recovery attacks or combine genome data in a specific way to breach existing private release assurances [23].

The decreased computational performance and restricted computational resource availability are other limiting factors of current cryptographic-based mechanisms. Even though several promising works have shown the feasibility of conducting federated GWAS leveraging multi-key HE [12] and SMC schemes [6], cryptographic methods exhibit extra communication costs (the case of SMC approaches) and demand increased storage resources (the case of HE-based solutions), for example. Similarly, Intel SGX, the most popular TEE-based technology, carries limitations regarding memory availability (only 96 MB is usable inside the processor's isolated – cryptographically protected – regions). In addition, recent works have shown that Intel SGX is vulnerable to side-channel attacks [21]. Such limitations oblige the deployment of data-oblivious versions of the privacy-protecting

algorithms, which decreases the overall performance of the solutions due to increased running time entailed by oblivious operations [27, 21, 2].

As a summary, despite enabling secure and privacy-preserving *processing* (e.g., by the adoption of cryptographic schemes), solutions for federated GWAS should also apply proper privacy-preserving *releasing* mechanisms since released results can be subject to genomic privacy attacks. As a result, impeding genomic privacy leaks throughout the entire pipeline of the study. Hence, we claim that an indispensable feature of federated GWAS solutions is to reconcile privacy-preserving *processing* with privacy-preserving *releasing*. Thus, creating an end-to-end privacy-aware collaborative environment.

2.2 Future research directions

In our previous work, Dynamic, Private and Secure GWAS (DyPS) [23] published at PETS'21, we addressed some of these challenges, namely (i) reconciling privacy-preserving processing (leveraging a TEE-enabled architecture) with private releases of federated GWAS (selecting genome data that can safely participate in GWAS releases that accommodate the conditions to protect releases against recovery (selecting data to keep the complexity of recovery attacks large enough) and membership attacks (reverse engineering genome-oriented statistical inference tests); (ii) extend the approach to allow safe and dynamic updates of GWAS statistics while enabling consent withdrawal and addition of new donors at any time; and at the same time (iii) supporting the presence of colluding federations members. We have leveraged Intel SGX as our TEE enabler. Our choice for SGX over other TEEs is arbitrary and following its adoption in previous works and by the increasing availability of SGX on cloud services [7, 19]. Nevertheless, our privacy-protecting mechanism applies well to other TEE implementations.

Nevertheless, there are still some open issues regarding several aspects debated in the previous section. For instance, studying the privacy vs. data utility trade-off tailored with DP-based mechanisms to certify that released statistics do not become impractical or overly inaccurate, which might compromise correct conclusions from GWAS results. In addition to that, there is a lack of mechanisms capable of determining safe conditions for safe GWAS releases provided adversaries might observe the presence of overlapping genome data. In particular, this interdependent threat model directly impacts the privacy guarantees enabled by existing solutions. Copying with this adversarial model is crucial since we anticipate that it will become common to detect particular genomes participating in several studies simultaneously. As one could expect, this novel interdependent privacy assumption would impact genomic privacy assurances in a crossed-over manner.

On the system side aspects, recent works have shown that multiparty cryptographic schemes can reasonably accommodate current federated GWAS security requirements. However, existing works still fail to offer holistic solutions, i.e., a workflow that simultaneously enables both privacy-preserving *processing* and *releasing* of GWAS.

Regarding TEE-based solutions, although Intel SGX is becoming deprecated on desktop versions due to side-channel attacks, Intel will continue manufacturing and providing support for Intel SGX on server platforms[1]. Unfortunately, this can limit the use of the technology mainly in distributed settings, e.g., by systems that leverage multiple TEE instances. Nevertheless, since Intel SGX will not be fully deprecated and because of the existence of other TEE solutions (e.g., ARM TrustZone), we envisage as future improvements the creation of built-in oblivious operations (i.e., applied on the hardware level) in the subsequent versions of TEE-based technologies.

As we can observe, paving the way to enable secure and end-to-end private workflows for federated GWAS is in its early stages and therefore requires not only scalable but also usable, fully privacy-aware solutions. Indeed, as identified in previous sections, most of these challenges are inherited from the existence of interdependent and multi-party genomic privacy [17, 15].

Although the community has been making great efforts to develop fully privacy-aware solutions, we claim that the field still requires a standardization of genomic privacy vs. data utility metrics. Therefore, we envision as another future direction the creation of studies to categorize and combine privacy guarantee parameters with federated GWAS (system) settings. Namely, assessing reasonable values of ϵ for DP-based approaches or acceptable confidence levels to be specified in statistical inference analyses while combining with system settings such as the desired number of dynamic releases a study will sustain, the number of honest and supported colluding members the federation will support, removal/addition rate of genomes, etc. This would enable not only the general community but also GWAS federations to evaluate and decide the most suitable approach according to their expectations.

3. CONCLUSION

In this work, we present an overview and remaining challenges to support the creation of end-to-end private and practical federated GWAS. In particular, we show that many aspects of interdependent and multi-party genomic privacy are still uncovered and therefore require suitable treatment. Even though previous works have shown that it is already possible to (i) enforce several privacy guarantees to protect the data of federation members whose genome datasets are to be outsourced for federated analysis, and (ii) measure personal genomic privacy (i.e., enforcing privacy on a per-genome basis) while updating GWAS statistics, when studies start overlapping and/or more dependent data become present (e.g., the presence of individuals' relatives), new privacy risks are still to be quantified and suitable mitigation implemented. In addition, we highlight the need for usable solutions capable of complying with existing data privacy regulation guidelines, which is a key factor in determining the success and large-scale adoption of federated GWAS in the future.

[1] https://github.com/intel/linux-sgx/issues/760

4. REFERENCES

[1] N. Almadhoun, E. Ayday, and Ö. Ulusoy. Differential privacy under dependent tuples—the case of genomic privacy. *Bioinformatics*, 36(6):1696–1703, 2020.

[2] A. Asvadishirehjini, M. Kantarcioglu, and B. Malin. A framework for privacy-preserving genomic data analysis using trusted execution environments. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 138–147. IEEE, 2020.

[3] E. Ayday and M. Humbert. Inference attacks against kin genomic privacy. *IEEE Security & Privacy*, 15(5):29–37, 2017.

[4] M. M. A. Aziz, S. Kamali, N. Mohammed, and X. Jiang. Online algorithm for differentially private genome-wide association studies. *ACM Transactions on Computing for Healthcare*, 2(2):1–27, 2021.

[5] M. Backes, P. Berrang, M. Humbert, X. Shen, and V. Wolf. Simulating the large-scale erosion of genomic privacy over time. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1405–1412, 2018.

[6] D. Bogdanov, L. Kamm, S. Laur, and V. Sokk. Implementation and evaluation of an algorithm for cryptographically private principal component analysis on genomic data. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1427–1432, 2018.

[7] A. Bomai, M. S. Aldeen, and C. Zhao. Privacy-preserving gwas computation on outsourced data encrypted under multiple keys through hybrid system. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 683–691. IEEE, 2020.

[8] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1655–1658, 2018.

[9] J. Decouchant, M. Fernandes, M. Völp, F. M. Couto, and P. Esteves-Verissimo. Accurate filtering of privacy-sensitive information in raw genomic data. *Journal of biomedical informatics*, 82:1–12, 2018.

[10] I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, and E. Ayday. An inference attack on genomic data using kinship, complex correlations, and phenotype information. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(4):1333–1343, 2017.

[11] M. Fernandes, J. Decouchant, M. Völp, F. M. Couto, and P. Esteves-Verissimo. Dna-seal: sensitivity levels to optimize the performance of privacy-preserving dna alignment. *IEEE Journal of Biomedical and Health Informatics*, 24(3):907–915, 2019.

[12] D. Froelicher, J. R. Troncoso-Pastoriza, J. L. Raisaro, M. A. Cuendet, J. S. Sousa, H. Cho, B. Berger, J. Fellay, and J.-P. Hubaux. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nature communications*, 12(1):1–10, 2021.

[13] M. Herper. *Illumina Promises To Sequence Human Genome For $100 – But Not Quite Yet.* https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet. Accessed on: January 7th, 2022.

[14] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8), 2008.

[15] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):1–31, 2017.

[16] M. Humbert, D. Dupertuis, M. Cherubini, and K. Huguenin. Kgp meter: Communicating kin genomic privacy to the masses. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, page 20, 2022.

[17] M. Humbert, B. Trubert, and K. Huguenin. A survey on interdependent privacy. *ACM Computing Surveys (CSUR)*, 52(6):1–40, 2019.

[18] H. K. Im, E. R. Gamazon, D. L. Nicolae, and N. J. Cox. On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics*, 90(4):591–598, 2012.

[19] C. Kockan, K. Zhu, N. Dokmai, N. Karpov, M. O. Külekci, D. P. Woodruff, and S. C. Sahinalp. Sketching algorithms for genomic data analysis and querying in a secure enclave. In *RECOMB*, pages 302–304. Springer, 2019.

[20] C. Lambert, M. Fernandes, J. Decouchant, and P. Esteves-Verissimo. Maskal: Privacy preserving masked reads alignment using intel sgx. In *2018 IEEE 37th Symposium on Reliable Distributed Systems (SRDS)*, pages 113–122. IEEE, 2018.

[21] A. Mandal, J. C. Mitchell, H. Montgomery, and A. Roy. Data oblivious genome variants search on intel sgx. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Springer, 2018.

[22] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[23] T. Pascoal, J. Decouchant, A. Boutet, and P. Esteves-Verissimo. Dyps: Dynamic, private and secure gwas. *Proceedings on Privacy Enhancing Technologies*, 2:214–234, 2021.

[24] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, and J.-P. Hubaux. Medco: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(4):1328–1341, 2018.

[25] M. N. Sadat, M. M. Al Aziz, N. Mohammed, F. Chen, X. Jiang, and S. Wang. Safety: secure gwas in federated environment through a hybrid solution. *TCBB*, 16(1):93–102, 2018.

[26] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. Genomic privacy and limits of individual

detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[27] E. Stefanov, M. V. Dijk, E. Shi, T.-H. H. Chan, C. Fletcher, L. Ren, X. Yu, and S. Devadas. Path oram: An extremely simple oblivious ram protocol. *Journal of the ACM (JACM)*, 65(4):1–26, 2018.

[28] O. Tkachenko, C. Weinert, T. Schneider, and K. Hamacher. Large-scale privacy-preserving statistical computations for distributed genome-wide association studies. In *Asia CCS*, 2018.

[29] N. Von Thenen, E. Ayday, and A. E. Cicek. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics*, 35(3):365–371, 2019.

[30] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009.

[31] E. A. Zerhouni and E. G. Nabel. Protecting aggregate genomic data. *Science*, 322(5898):44–44, 2008.

[32] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang. To release or not to release: Evaluating information leaks in aggregate human-genome data. In *Esorics*, 2011.