2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) | 979-8-3503-2385-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/TPS-ISA58951.2023.00011

Ensuring Trust in Genomics Research

Erman Ayday
Department of Computer and Data Sciences
Case Western Reserve University
Cleveland, OH
exa208@case.edu

Jaideep Vaidya
Management Science and Information Systems Department
Rutgers University
Newark, NJ
jsvaidya@business.rutgers.edu

Xiaoqian Jiang
Department of Data Science and Artificial Intelligence
University of Texas - Health
Houston, TX
xiaoqian.Jiang@uth.tmc.edu

Amalio Telenti
Dept. of Integrative Structural and Computational Biology
Scripps Institute
La Jolla, CA
atelenti@scripps.edu

***Abstract*— Reproducibility, transparency, representation, and privacy underpin the trust on genomics research in general and genome-wide association studies (GWAS) in particular. Concerns about these issues can be mitigated by technologies that address privacy protection, quality control, and verifiability of GWAS. However, many of the existing technological solutions have been developed in isolation and may address one aspect of reproducibility, transparency, representation, and privacy of GWAS while unknowingly impacting other aspects. As a consequence, the current patchwork of technological tools only partially and in an overlapping manner address issues with GWAS, sometimes even creating more problems. This paper addresses the progress in a field that creates technological solutions that augment the acceptance and security of population genetic analyses. The text identifies areas that are falling behind in technical implementation or where there is insufficient research. We make the case that a full understanding of the different GWAS settings, technological tools and new research directions can holistically address the requirements for the acceptance of GWAS.**

Keywords—genomics research, trust, privacy, transparency, representation

I. INTRODUCTION

Genome Wide Association Studies (GWAS) scan the genomes of thousands of individuals to identify genetic markers that are associated with a trait or a disease. The output of a GWAS consists of statistics such as p-values, chi-square values, and odds ratios. These outputs and their derivatives can be used for various purposes, including (i) identifying gene/variant and phenotype correlations, (ii) generating data to build polygenic risk scores (PRS) for prediction and causal inference, and (iii) learning more broadly about the biology of a trait. Today, GWAS are conducted in different settings, including (i) local setting, in which a researcher conducts the study on a local dataset and shares the result (e.g., through research papers), and (ii) collaborative setting, in which two or more researchers combine their datasets for joint research. Alternatively, when data cannot be shared, statistics support collaborative meta-analysis over multiple independent studies.

Regardless of the setting, key requirements for trust in GWAS are reproducibility, transparency, representation, and privacy (herein referred to as "challenges of GWAS"). Technical reproducibility enables the validation of the research and its further development by other researchers. Transparency ensures that the pipeline of data collection, analysis, and dissemination is open and accessible to all stakeholders. A balanced representation ensures that all populations can benefit from the research, which requires the participation of people from different backgrounds (especially underrepresented populations) in research studies. Privacy is threatened by loss of anonymity. Note however that "trust" in genomic research also depends on the participating parties and the nature of the study. For a given study, some of the aforementioned key requirements may be directly required to develop trust for some parties, white others may be indirect requirements.

To address these key requirements, technology can come to the rescue, by providing assurance for all of these challenges of GWAS. For example, quality control tools mitigate bias from the analysis by eliminating low-quality, noisy, or incomplete data and also ensure the robustness of the analysis. Similarly, privacy-enhancing technologies guarantee that collected data is processed and results are shared in a privacy-preserving way. Verifiability tools help identify potential miscalculations during the GWAS computation. However, many technological solutions have been developed in isolation and in different contexts, thus partially and in an overlapping manner solving individual challenges of GWAS. For example, technical solutions for privacy-preserving GWAS have been extensively studied [1–3]. However, implementing such privacy-preserving solutions in isolation is not sufficient to address all key challenges of GWAS. Furthermore, while they may address one challenge of GWAS, they may unknowingly compromise other aspects.

Here, we only focus on reproducibility, transparency, representation, and privacy because these values are essential to the scientific process. However, these aspects do not cover the full spectrum of ethically, socially and legally relevant challenges in GWAS. In particular, this paper does not cover privacy-related conditions for sharing DNA samples, or privacy issues linked to communication of findings and results to research participants. Note that there are other requirements for

979-8-3503-2385-6/23/$31.00 ©2023 IEEE
DOI 10.1109/TPS-ISA58951.2023.00011

the acceptance of GWAS studies, such as the misuse of findings. Researchers must take care to ensure that the results of a study are not used in a way that could harm or exploit vulnerable populations. Additionally, they must work to ensure that any benefits of the study are shared equitably among all participants.

In this work, we first discuss the key challenges of GWAS. Next, we summarize existing technical solutions for responsible management of genomic data. Finally, considering different GWAS settings, we discuss how additional technical challenges need to be addressed to pave the way toward greater trust in GWAS. Rather than just comprehensively surveying all existing technical solutions, which has been the object of recent work [1–3], the goal of this paper is to identify and prioritize different requirements, and then work to resolve any conflicts between them to develop technical solutions for responsible GWAS. This process can help ensure that the final technical solution meets the needs of all stakeholders.

The reader should also use this work to address the next challenges emerging from a broader use of whole exome and genome sequencing technologies. The experience with GWAS should guide the assessment of reproducibility, transparency, representation, and privacy unique to sequence data.

II. CHALLENGES OF GWAS

A. Technical reproducibility of results

In this paper, we define reproducibility as obtaining consistent results using the same data and code as the original study. Reproducibility of research results is crucial to validate existing research and build on top of it. To provide reproducibility, researchers typically share their workflows (methodologies). Such workflows support tracking research data through preprocessing, analyses, and interpretation. Currently, workflows are stored as in [4], where all the steps are necessary to go from the initial input(s) to the final output(s). For instance, if research is published, it should be possible to have access to the research findings (workflow output), their associated metadata (e.g., model parameters, demographics of the research participants, and assumptions), and the input dataset. There exists a number of tools, such as Taverna [5], Kepler [6], and VisTrails [7] for building workflows and capturing provenance on them, which facilitates reproducibility and interpretation of the research results.

To address the reproducibility requirements, verifiability tools (as in Section III.C) can be utilized by formulating reproducibility as the verifiability of research findings. On the other hand, providing the input dataset raises privacy concerns and most of the time, sharing it with other parties is subject to a complex institutional review board (IRB) process. Therefore, there is a need for solutions that allow reproducibility in a privacy-preserving and practical way (as discussed in Section IV.A.2).

In this paper, we primarily focus on the issues surrounding data accessibility and sharing as major obstacles to reproducibility. While we acknowledge that technical obstacles, such as maintaining functional installation scripts and changing paths, can indeed pose challenges to reproducibility, a detailed discussion on these aspects is beyond the scope of this article. However, it is worth mentioning that utilizing containerization technologies, such as DockerHub, and repositories like Figshare and Zenodo can help address some of these technical challenges and promote reproducibility. These solutions require a certain level of technical expertise, but can significantly improve the ease of reproducing research workflows in the long run.

B. Transparency

In this paper, we define transparency as ensuring that the pipeline of data collection, analysis, and dissemination is open and accessible to all stakeholders. Research participants would like to know the consequences of sharing their data both in terms of risk and benefit. As discussed below, the primary privacy risks of sensitive data sharing are re-identification [8,9], attribute inference [10], and membership inference [11]. Depending on the GWAS setting, type of collected data, and how the research outcome is shared, such privacy risks should be communicated to the research participants to provide transparency.

To address the transparency requirement, using the privacy risk quantification tools (in Section III.A.5) and conveying the outcome to the research participants will help (i) inform participants about the consequences of their shared data, and hence provide transparency while data sharing and (ii) provide data minimization considering the required usage of the data and preserving privacy against the identified vulnerabilities. Thanks to such tools and algorithms, research participants are made aware of their privacy risk/utility tradeoff when sharing data which will allow them to provide fully informed consent for the collection and use of their data. From the researchers' point of view, it is important to understand the incentives of the research participants to take part in the research activities. In addition to knowing the potential privacy consequences, participants also would like to know other factors before they decide to take part in research studies. Such factors may include credibility of the research institution, benefit of the research study for the participants and for the population at large, interdependent privacy risks (privacy risks that may occur for the family members of the research participants due to the data shared data by the participants), the duration of data storage by the research institution, and data use agreements (e.g., possible commercial use). Such incentives of participation can be formalized and analyzed using game theoretical formulations as in [12].

C. Representation

In this paper, we define representation as ensuring that all populations can benefit from the research. There is unequal representation of human populations in genomic research [13]. Based on [14], data from individuals of European ancestry account for at least 78% of GWAS while individuals of African ancestry account for 2.4%. In addition, over 70% of samples included in GWAS originate from only three countries (USA, Iceland, and UK) [14]. On the other hand, associations between genetic variants and traits may not be accurate unless the study involves individuals belonging to diverse ancestral backgrounds. Models developed using non-diverse populations may not be generalizable, and treatments developed based on results of association studies on non-diverse populations may be ineffective in underrepresented groups [15–18].

To address these issues, it is crucial to increase the participation of underrepresented populations in genomic research. One barrier for participation of underrepresented populations in research studies is "trust". There are technical solutions that enable trust in GWAS, in particular trust in privacy (next Section II.D). These include privacy-enhancing technologies and tools that provide privacy risk quantification to clearly show the consequences of data sharing to the research participants (as in Section III.A). In addition, researchers, using the quality control tools (e.g., population stratification, as discussed in Section III.B), can better understand the backgrounds of the research participants to mitigate bias. More generally, by knowing how data is collected, processed, and shared, individuals may be more willing to participate. However, it is not clear that risk quantification tools will advance trust in GWAS, and it is not obvious that this would in-turn promote participation in GWAS studies from more diverse human populations.

We recognize that there are unaddressed technical requirements related to representation, as well as active research and development in this domain. Nevertheless, our article's objective is to underscore the interconnection between these concerns and the core themes of reproducibility, transparency, and privacy, rather than delivering an extensive analysis of the representation challenge. Examining the link between representation and our central themes will bolster the comprehension of the hurdles and potential solutions in data-driven research.

D. Privacy

In this paper, we define privacy as ensuring that the research participants remain safe against known inference attacks. Privacy is a pressing challenge in conducting GWAS [1–3]. The worst-case scenarios that may occur due to misuse of genomic data (in the case of lack of privacy) include genetic discrimination in health/life insurance, employment, education, etc. and blackmail (e.g., considering unknown paternity situations). GWAS data (i.e., set of SNPs) can be used to infer polygenic risk of diseases of individuals or for paternity cases. This is an example of attribute inference attack (in Section 3.1.5). On the other hand, membership inference attack (in Section III.A.5) is typically more serious. In membership inference, using GWAS outcome, an attacker can infer the membership of a victim to the case group that is used in the study. Case group typically has a label corresponding to the trait under study (e.g., Cancer, Parkinsons, Autism, HIV, etc.). Therefore, inferring the membership of a victim to the case group is equivalent to inferring the corresponding label for the victim. Lack of privacy in the system may cause betrayal or neglect of GWAS participants' trust. This may lead to the participants not donating their data to research, and hence pose a barrier for the advancement of the field.

Most existing work considers privacy in four stages in the process: (i) privacy during data collection from the research participants, (ii) privacy during data pre-processing (i.e., while identifying and removing certain data records or attributes from the research study), (iii) privacy during GWAS computation (in *outsourced* and collaborative settings), and (iv) privacy while sharing the research outcomes (Figure 1). In addition, we also consider privacy risk quantification tools and identification of privacy requirements. Methods that quantify the privacy risk can provide transparency to individuals and organizations about the risks associated with sharing genomic data. The methods presented below in Section III.A can help individuals and organizations make informed decisions about whether to share genomic data and can help identify and address privacy risks. Quantifying the privacy risk can also help organizations improve their privacy practices.

III. Technical Solutions to address Challenges in GWAS

The technical approaches and existing solutions to address challenges in GWAS are depicted in Figure 1. Those techniques address privacy protection, quality control, and solutions for verifiability. Note that one common natural requirement for all settings is security, which includes keeping data encrypted at rest, secure communication between the involved parties, and access control to datasets and research results. Such security requirements can be achieved by using existing tools, and hence they are not discussed in this paper.

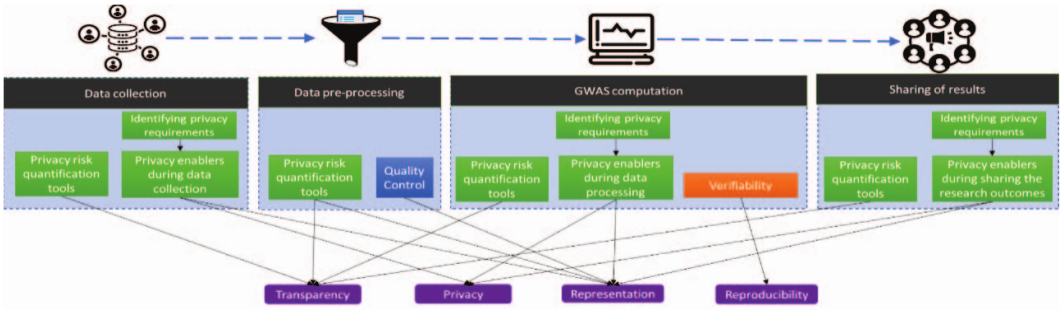

Figure 1: Technical solutions. The boxes in different GWAS steps represent the technical tools that are required during the corresponding step: Green boxes represent privacy enablers, dark blue box represent quality control, and orange box represent verifiability. Privacy requirements (e.g., via standards, policies, or law) provide input for the privacy enablers at each GWAS step. Purple boxes at the bottom represent the key challenges, and the arrows between technical tools and the challenges show which technical solution helps address a given challenge.

A. Privacy protection solutions

As mentioned before, this paper is not meant to be an extensive survey of privacy-preserving solutions for GWAS, but rather to identify key aspects of privacy requirements and the corresponding challenges for GWAS. Thus, in the following, we focus on specific solutions that have been proposed to protect the privacy of genomic data in the life cycle of GWAS.

1) Privacy during data collection. Generally, participants' data is collected to conduct GWAS under consent and data use agreements without any obfuscation. However, it is possible to add privacy guarantees to data collection. Privacy guarantees help in case the research dataset is breached due to vulnerabilities in the system. To share a participant's data with a researcher, one promising direction is to utilize local differential privacy (LDP) techniques.

LDP [19,20], a variant of differential privacy [21] (which will be discussed in Section III.A.4), is a state-of-the-art model to preserve the privacy where participants perturb (randomize) their inputs before sharing with researchers, and this provides indistinguishability guarantees for the shared data. Such guarantees help provide anonymization, and hence LDP has the potential to fulfill legal de-identification requirements in HIPAA or the GDPR. After receiving data from the participants under LDP, the data aggregator (researcher) applies estimation techniques to partially eliminate noise before processing the collected data. Collecting perturbed data from more individuals decreases the accuracy loss due to randomization. Hence, practical usage of LDP-based techniques needs a high number of research participants, which limits the utility of LDP-based techniques. To overcome the accuracy loss due to LDP, a s shuffling technique has been proposed [22,23], wherein a trusted party receives the data from participants and permutes them before sending them to the researcher. Another approach to improve the utility of LDP is providing different privacy protection for different inputs [24,25]. Another issue with using LDP-based techniques for genomic data collection is the correlations in genomic data (e.g., linkage disequilibrium). In previous work [26], we showed the privacy risks when the shared data includes correlations, and we explored a variant of LDP that considers correlations in the data and optimizes utility without compromising privacy.

Another privacy challenge during data collection arises due to correlations between the genomes of people from the same family (e.g., interdependent privacy issues). Two main directions have been proposed: (i) optimization-based schemes [27], in which a research participant shares their genome by both considering privacy budgets of their family members (i.e., privacy preferences, which represent how much individuals want to reveal about their genomic data) and the utility of the shared data; and (ii) selective sharing techniques [28], which aim to protect sensitive parts of genomic data both for research participants and their family members while maximizing the utility of the shared data. Considering that people of the same family might have very different opinions about how to protect and whether or not to reveal their genome, genomic data sharing by family members can also be studied in a game-theoretic setting [12], where a closed-form Nash equilibria can be defined in different settings and the game evolution analyzed when relatives behave altruistically.

2) Privacy during data pre-processing. The techniques in this category mainly focus on quality control (discussed in Section III.B). Cho et al. [29] considered conducting the quality control steps in an outsourced environment in a privacy-preserving way using secure multiparty computation. Huang et al. [30] considered privacy-preserving execution of a limited number of quality control steps before meta-analysis, with the goal of approving or rejecting a collaborative study. In Section IV.A.1, we will argue that existing privacy-preserving quality control tools are insufficient, especially for collaborative GWAS settings, and we will highlight the additional privacy requirements during data pre-processing.

3) Privacy during GWAS computation. In the settings of outsourced and collaborative research, the existing privacy-preserving GWAS computation techniques can be categorized into three main categories: (i) techniques that rely on homomorphic encryption [31–33], in which the computation is done at an untrusted cloud or among the participants over

encrypted data; (ii) techniques that utilize secure multiparty computation [29,32,34,35], in which research dataset is distributed to several non-colluding servers and the computation is done among such servers; and (iii) techniques that utilize secure hardware [36], in which the computation is done inside a secure, tamper-proof enclave. For the last category, trust in hardware manufacturers is a crucial aspect for any computation that relies on hardware, as the reliability and security of the hardware directly impacts the overall performance and safety of the system. In the context of privacy-preserving GWAS, this trust is particularly important, because of the handling of sensitive data and computations. The actual concern in this scenario should be focused on the attestation procedure, which is the process of verifying the integrity and authenticity of the hardware being used. A robust attestation procedure can help mitigate potential risks and ensure that the hardware is secure and trustworthy. Additionally, attention should be given to potential side-channel leaks, which are vulnerabilities that could be exploited to gain unauthorized access to sensitive data or computations.

The main goal of all these techniques is to conduct GWAS computations without revealing the research dataset to the cloud server (in the outsourced setting) or to other collaborating researchers (in the collaborative setting). Among the aforementioned techniques, homomorphic encryption and secure hardware-based techniques enable collaborative privacy-preserving GWAS. However, techniques that rely on homomorphic encryption typically suffer from scalability problems, whereas techniques that rely on secure hardware suffer due to potential side-channel attacks and trust issues (as they require a level of trust on the hardware manufacturer). Alternatively, secure multiparty computation-based techniques enable outsourcing the computation of GWAS to third-party servers (which typically have high computational power). However, the real-life implementation of secure multiparty computation-based techniques is non-trivial, as they require the existence of several non-colluding servers in the system.

4) Privacy while sharing the research outcomes. Most institutions, including the US National Institutes of Health (NIH), allow sharing of GWAS results [37]. However, it has been shown that sharing of aggregate statistics may lead to privacy risks for the dataset participants [11]. To reduce this risk, some researchers have proposed using the model of differential privacy to mitigate membership inference attacks when releasing summary statistics. Differential privacy [21] is a concept to preserve the privacy of records in statistical databases while publishing statistical information about the database. Fienberg et al. used the differential privacy concept for sharing statistics, such as minor allele frequencies and chi-square values [38]. Yu et al. extended this work and presented a scalable algorithm for any arbitrary number of SNPs [39]. Johnson and Shmatikov proposed using a variation of differential privacy for the computation and release of statistics about a genomic database [40]. Tramer et al. also studied the tradeoff between privacy and utility provided by differential privacy [41]. Although differential privacy provides strong privacy guarantees, noise added to GWAS statistics to achieve differential privacy results in significant degradation in the correctness of the research outcome.

5) Privacy risk quantification tools. Many researchers have worked on identification and quantification of privacy risk due to re-identification, attribute inference, and membership inference attack. Re-identification (deanonymization) links anonymous data to contributing subjects. Attribute inference aims to infer missing (hidden) attributes of an individual from the observed ones. Membership inference aims to infer the membership of an individual in a private database using the results of the queries that are obtained from this database. Today, it is possible to quantify (i) re-identification risk in an anonymized genomic dataset depending on varying amounts of auxiliary information [8], (ii) membership inference risk due to sharing of aggregate statistics about a genomic dataset [11], (iii) membership inference risk due to genomic data sharing beacons [42,43], (iv) attribute inference risk due to genotype-phenotype correlations [9,44], and (v) interdependent privacy risks (e.g., decrease in genomic privacy of an individual due to sharing by a family member) [10].

Attribute inference attacks are typically modeled and quantified via inference algorithms that consider statistical correlations between hidden (sensitive) data points and auxiliary information. Membership inference attacks are typically modeled using a likelihood ratio test to quantify the power of an attacker. Using such tools developed in previous work [9,10,42–44], it is possible to quantify the risk of attribute inference on the data collected from the research participants. Furthermore, to quantify the privacy risk when researchers share aggregate statistics, tools that quantify the membership inference risk can be used.

6) Identifying privacy requirements. Considering all different privacy-preserving solutions we have discussed, one open research question that remains is: how much privacy is enough? Or equivalently, can we accept the data sharing policies of some institutions as baseline for privacy? For instance, using the aforementioned privacy risk quantification tools (in Section III.A.5), NIH genomic data sharing policy can be represented quantitatively and accepted as a baseline privacy requirement. Such a quantifiable privacy baseline would also help in selecting the privacy parameters (e.g. the epsilon parameter in differential privacy) or how much data/metadata to share in outsourced and collaborative GWAS settings.

B. Quality control

It is crucial to assess the quality of the datasets that GWAS is conducted on to achieve reliable genetic associations between

traits and diseases. If the research datasets are not properly curated (e.g., if dependent/correlated data records are not removed or heterogeneous populations are not controlled for), the summary statistics calculated as a result of GWAS may lead to biased associations. Furthermore, studies that implement wrong statistics (obtained via GWAS without quality control) are unlikely to provide reproducible results [45].

Previous work investigated the implementation of quality control procedures in local GWAS datasets with the aim of creating high-quality datasets that will result in accurate research outcomes [15–18]. The quality control steps which are extensively used include eliminating samples or markers with a significant number of missing values in the dataset, identifying and eliminating sex inconsistency, eliminating variants with low minor allele frequency, eliminating the variants that are not compliant with Hardy-Weinberg equilibrium, eliminating the relatedness between the samples (e.g., removing samples with high kinship relationships), and population stratification (e.g., eliminating or correcting for heterogeneous populations).

Conducting such steps for non-collaborative research studies is straightforward, since both quality control and the research are done locally by the same researcher. On the other hand, in collaborative settings, although some of the simple quality control steps can be done locally by each researcher, some of the steps have to be conducted on the combined dataset and may impact requirements such as privacy. For instance, sample missingness can be dealt with at each research site independently and does not need to be repeated in the combined dataset. However, relatedness or population stratification has to be done on the combined dataset. Two or more datasets that include completely homogeneous populations and no family members, when looked at in isolation, may turn out to contain bias when they are combined, thus impacting diversity and inclusion (e.g., considering genealogy) [46]. For genomic studies, it is desirable to identify the ancestry of the research participants from their genomic data (e.g., via a principal component analysis). In addition, for collaborative studies (e.g., meta-analysis), even a perfect protocol cannot fully compensate for not having access to individual participant data, which would guarantee standardized quality control [47]. This clearly shows the importance of quality over the federated data. In addition, as we will discuss in Section IV.A.1, direct use of quality control tools may result in privacy concerns.

C. *Verifiability*

It is crucial to verify the correctness of published research. Computational errors might occur during the workflow (e.g., the published results/statistics or the metadata may be computed wrong) or during quality control (e.g., a researcher might use low-quality data to conduct the research). It is trivial to verify the correctness of the research findings if, besides the workflow and its associated metadata, the input dataset is provided. In most applications, provenance (origin or the data or computation) is captured using these components. However, the input dataset might not be released as it may contain sensitive information about individuals (e.g., personal records). In such cases, verifying the correctness of the computations becomes non-trivial.

There exist several works in the field of verifiable computation, which aim to do various computations on the cloud while verifying the correctness of the returned results [48,49]. However, the correctness of GWAS results cannot be easily verified by those general tools without having access to the original research dataset. One alternative may be to use homomorphic authenticators [50], but they are impractical for statistical analysis on large datasets due to the high computation burden. In addition, Zero-knowledge proof (ZKP), which is a cryptographic technique introduced by Shafi Goldwasser, Silvio Micali, and Charles Rackoff in the 1980s, allows one party to prove the validity of a statement without revealing any information about the statement itself, except for the fact that it is true. In the context of genomics research, ZKP can be used to authenticate the genomic data and results from genome computations, thus addressing the verifiability and transparency issues without directly sharing the original human genomic data for privacy protection. But, ZKP techniques are typically computationally expensive and may not scale well with the size of the genomic dataset. Designing efficient ZKP protocols for complex genomic computations is a non-trivial challenge, and may require a deep understanding of the underlying cryptographic primitives and genomic data structures.

Therefore, new research is needed to provide verifiability of GWAS computations. In addition, as we will discuss in Section IV.A.2, verifiability tools may also result in privacy concerns, and this should be taken into consideration when developing such tools.

IV. Conflicts for the Implementation of Technical Solutions in Various Settings

Conflicts arise when trying to provide technical solutions for several challenges of GWAS. For example, one potential conflict is between the need for privacy and the need for transparency of the research methodology. To protect the privacy of study participants, it may be necessary to keep certain data confidential. However, to ensure that the study is conducted in an open and transparent manner, it may be necessary to release some or all of that data. Identification of such conflicts is one of the main contributions of this article, and to the best of our knowledge, there is no previous work in that direction. We believe that identifying and resolving these conflicts will be valuable, and indeed crucial, for other researchers as they develop new technical solutions.

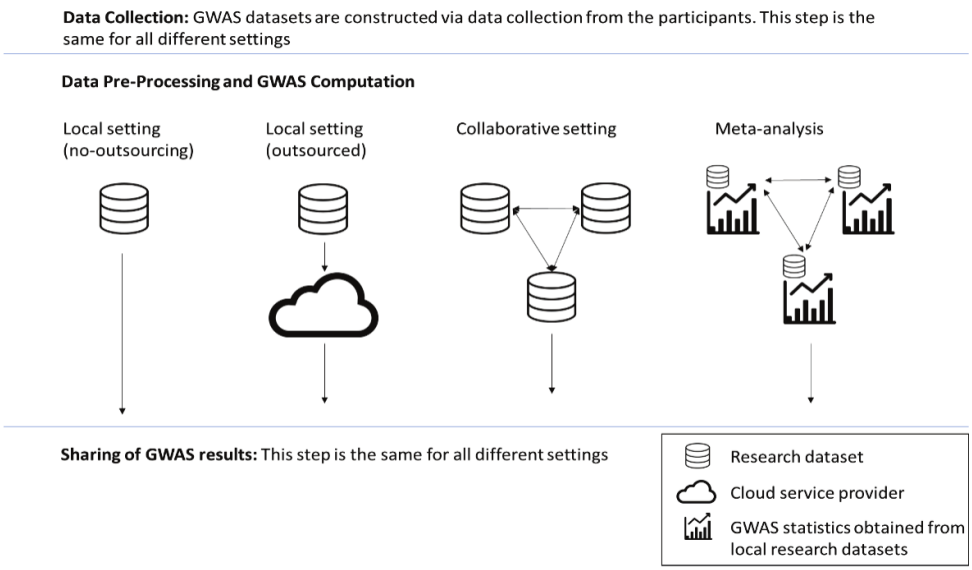

Figure 2: Pipeline for different GWAS settings.

A. Conflicts between the technical solutions

1) Quality control vs. privacy. As discussed in Section III.B, conducting quality control in the local setting (i.e., when a researcher conducts the research only by using their local dataset) does not create any privacy challenges since all quality control steps can be performed locally, without sharing data with a third party. On the other hand, in collaborative settings, researchers ideally want to consider the federated datasets to conduct the quality control steps due to potential bias that arise due to statistical relationships between the records across different datasets. However, sharing their datasets with each other (or with a centralized server) makes the dataset participants vulnerable to re-identification attacks (as discussed in Section III.A.5).

As discussed in Section III.A.2, some works consider privacy-preserving quality control in the local GWAS setting [29] or privacy-preserving execution of a limited number of quality control steps before meta-analysis [30]. However, existing works do not consider quality control steps that involve the federated datasets (e.g., relatedness or population stratification). To address this, existing cryptographic solutions (e.g., homomorphic encryption and its variants) can be used to check the kinship relationships between samples across different datasets, or to conduct principal component analysis on the federated dataset. However, such cryptographic solutions are not efficient, and they are not scalable for large-scale operations. Therefore, new privacy-preserving, IRB-compliant, and efficient techniques are needed to provide quality control, especially for collaborative GWAS.

In a recent work [51], considering the kinship elimination step of the quality control pipeline in a collaborative setting, we proposed a mechanism for 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. Similar mechanisms can also be developed to address other quality control steps that consider statistical correlations across data records (e.g., population stratification). Although such schemes do not require the collaborators to share their raw datasets with each other, they still require the exchange of a limited amount of metadata about the research datasets. As discussed in Section III.A.5, sharing of such data may lead to attribute inference and membership inference attacks, and the vulnerability of the proposed schemes against such attacks should be studied before using such schemes.

2) Verifiability vs. privacy. Achieving verifiability in isolation may result in privacy vulnerabilities, since most practical verifiability tools that provide reproducibility require providing the input dataset or significant information about the input dataset. However, as discussed in Section III.A.5, such information about the input dataset may result in re-identification risk for the participants, and hence it is not preferred/allowed by many institutions. Therefore, verifiability should be considered along with privacy. To achieve this, new cryptographic techniques can be developed using the existing verifiable computation tools [48,49] to allow researchers to assess the correctness of published research without having access to the original research dataset. Alternatively, as

discussed, efficient implementations of homomorphic authenticators [50] can serve as an alternative solution.

In a recent work [52], we also proposed a framework that verifies the correctness of the aggregate statistics obtained as a result of GWAS conducted by a researcher while protecting individuals' privacy in the dataset. In the proposed framework, the researcher keeps the dataset private while providing, as part of the metadata, a partially noisy dataset (that achieves local differential privacy). To check the correctness of the workflow output, the other researcher makes use of the workflow, its metadata, and the results of another GWAS (conducted using publicly available datasets) to distinguish between correct statistics and incorrect ones. Via evaluations using real genomic data, we show how the correctness of the workflow output (i.e., whether the output is computed correctly by the researcher) can be verified with high accuracy even when the aggregate statistics of a small number of variants are provided. We also quantify the privacy leakage due to the provided workflow and its associated metadata and show that the additional privacy risk (in terms of membership inference and re-identification) due to the provided metadata does not increase the existing privacy risk due to sharing of the research results (i.e., aggregate GWAS statistics, which is allowed by many institutions including the NIH [37]). Thus, our results show that the workflow output (i.e., research results) can be verified with high confidence in a privacy-preserving way. Such statistical solutions can also be a step towards providing provenance in a privacy-preserving way while providing guarantees to the users about the correctness of the results.

B. Addressing challenges in GWAS in different settings

We illustrate the pipeline for different GWAS settings in Figure 2. We illustrate the challenges of achieving the considered requirements in local and collaborative GWAS in Table 1. Also, a step by step guide with best practice approaches is presented in Figure 3.

1) Local (non-collaborative) GWAS. The research dataset is created by a single party (researcher or research site) as a result of data collection from the study participants. The research is completed without collaborating with external researchers and datasets. GWAS computation can be done either (i) directly at the research site (referred as "local GWAS computing") or (ii) it can be outsourced to a cloud service provider (referred as "outsourced GWAS computing").

Since local GWAS computing is done at the research site, achieving privacy is relatively easier and it has low overhead compared to the other settings. In terms of privacy risks, only membership inference or attribute inference are possible due to the shared GWAS results, which is a common risk for all settings. Privacy-preserving quality control of the research dataset is also less challenging in this setting, since the researcher can locally control the quality of the entire dataset before conducting GWAS. In contrast, achieving verifiability is challenging in the local setting, mainly due to the privacy requirements. To have both computation and results verified by external parties, the researcher needs to share information about the research dataset, which should be done by considering the privacy and liability requirements.

The case of outsourced GWAS computing implies that researchers may opt to benefit from the rich computational and storage capabilities of a cloud service to provide. Here, the computation at the cloud server should also be done in a privacy-preserving way using the techniques discussed in Section III.A. The main privacy risks are deanonymization or membership inference are possible due to the shared dataset. If the quality control is done before outsourcing the data to the cloud server, it is the same as the previous setting. However, if the quality control is done in the cloud, it becomes more challenging due to the privacy requirement. As discussed in [29], Cho et al. consider conducting the quality control steps in an outsourced environment in a privacy-preserving way using secure multiparty computation. There are challenges to comply with the requirements of privacy-preserving verifiability - similar to what was the case in the absence of outsourcing. As an additional step, the researcher may also verify the correctness of the server's computation.

Figure 3 illustrates the steps to conduct GWAS in a local setting: from data collection and pre-processing to the sharing of results (GWAS statistics).

2) Collaborative GWAS. For many diseases, the amount of data collected at any single site may be insufficient for a GWAS with high statistical power. Collaborative genomic data analysis is an essential tool that can unlock the potential of genomic data while minimizing the costs and other constraints associated with carrying out very large studies. In a collaborative GWAS setting, two or more researchers aim to conduct collaborative research and each researcher has its own research dataset. The research protocol requires IRB approval in order to exchange the research datasets. After combining the datasets, the scenario becomes similar to the local setting. Alternatively, researchers may outsource the computation of GWAS to a third party or may establish distributed computation directly among themselves. All these alternatives can be done under certain privacy guarantees, as discussed in Section III.A.

In general, achieving privacy in the collaborative setting is more challenging since each party contributes its own dataset with different privacy requirements. One trivial solution is to pool all research datasets in a cloud server and conduct GWAS on the federated dataset, however such an approach results in re-identification risk for the dataset participants, and hence it is not allowed by many institutions. In addition, in the collaborative setting, homomorphic encryption-based solutions

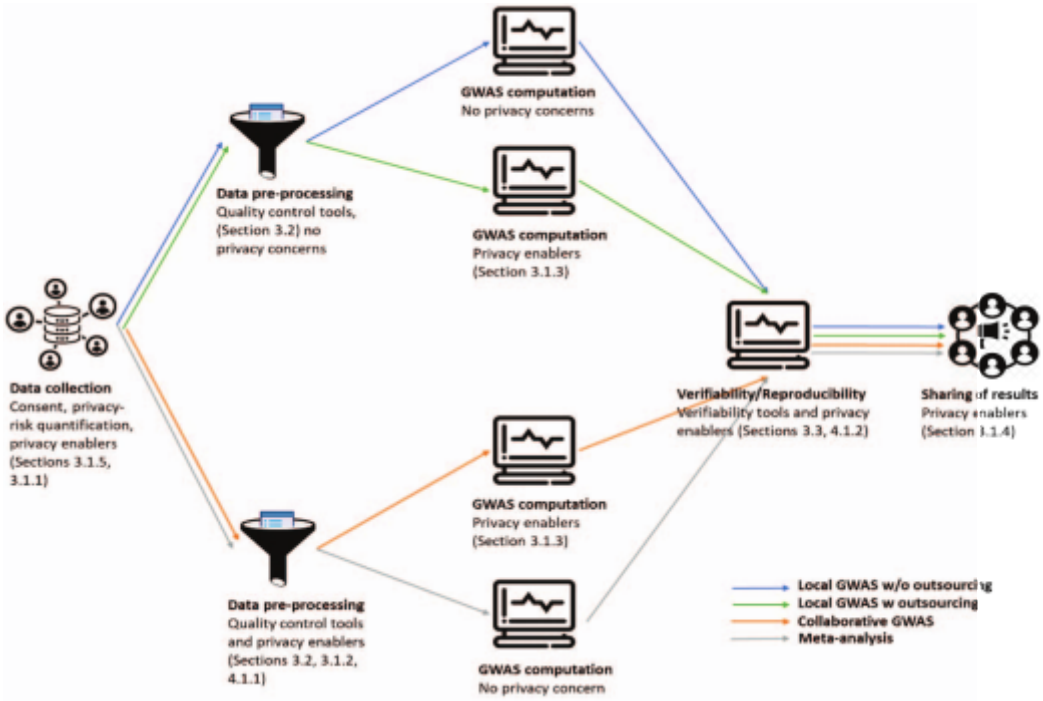

Figure 3: Guidelines for the researchers for the best practices while conducting GWAS under different settings.

typically do not scale well. De-anonymization or membership inference are possible due to the shared datasets.

Achieving quality control together with the privacy requirement in the collaborative setting is also more challenging compared to the local setting. New solutions, such as recent work [51], that addresses the kinship elimination step of the quality control pipeline in a privacy-preserving way, are required. Similarly, achieving privacy-preserving verifiability is more challenging than in the local setting since the collaborators first need to verify the correctness of each researcher's (or the cloud servers') computations first, and then they need to collaboratively provide information to other researchers for the verifiability of the entire computation. Doing these while also achieving the privacy requirements becomes non-trivial.

There is also the possibility to perform collaborative GWAS via meta-analysis. Meta-analysis is the statistical combination of results from separate studies. The researchers, after they conduct "local GWAS", share the locally computed results with each other and aim to derive the GWAS results that would have been obtained from the combined/federated dataset (without sharing the research datasets with each other). Such a setting benefits from collaborative research without taking high privacy risks (since parties do not share their research datasets with each other or with another third party). On the other hand, it has been shown that sharing of local results reveals information about the individual datasets, which is the main privacy risk for this setting. In particular, membership inference or attribute inference are possible due to the shared GWAS results by the collaborators, which can be alleviated using the tools introduced in Section III.A.5.

Achieving quality control in collaborative GWAS via meta-analysis is similar to the situation in the collaborative setting with dataset sharing. Quality control operations are done on the federated dataset, which makes it more challenging to fulfill the privacy requirements. New solutions, such as our recent work [51], which proposes a mechanism to identify correlated records across multiple data repositories in a privacy-preserving manner, are required. Verifiability in the setting of collaborative GWAS via meta-analysis requires that each researcher verify each other's local computations. The most challenging part is the verifiability of the global GWAS result. Similar to the collaborative setting with dataset sharing, our recent work [52], which proposes a technique to verify the correctness of the aggregate statistics obtained as a result of GWAS conducted by a researcher while protecting individuals' privacy in the dataset, can be an option.

GWAS setting	Privacy during GWAS computation	Privacy-preserving quality control	Privacy-preserving verifiability
Local without outsourcing	trivial	trivial	challenging
Local with outsourced	challenging	trivial	challenging
Collaborative via sharing datasets	challenging	challenging	challenging
Collaborative via meta-analysis	trivial	challenging	challenging

Table 1: Jointly addressing challenges in various settings.

We note that while the meta-analysis approach is comparatively simple to implement (compared to other privacy-preserving collaborative GWAS solutions), it has many limitations. It can be difficult to assess the quality of the studies included in a meta-analysis, leading to unreliable results. Additionally, if the studies included in the meta-analysis are not representative of the population of interest, the results may be biased. Meta-analysis also relies heavily on the accuracy and completeness of the data from the various studies, and it can be difficult to identify and control for potential confounding factors. Recently, privacy concerns have also been identified with meta-analysis [11]. Clearly, better solutions are necessary.
Figure 3 illustrates the steps to conduct GWAS in a collaborative setting: from data collection and pre-processing to the sharing of results (GWAS statistics).

C. Addressing challenges in GWAS for different parties

Trust in genomic research can be understood differently for the general public and the researchers. Here, general public can include (i) data owners whose data is used to conduct the research and (ii) patients who potentially benefit from the findings of the research. On the other hand, researchers may be (i) individuals or collaborators who conduct the research and (ii) individuals who use the results of the research. Thus, "trust" and the set of key requirements that need to be met depends on the parties and the nature of the genomic study. We discuss the direct and indirect requirements to establish for each involved party in the following paragraphs.

Data owners. Such individuals share/donate their data to construct the research datasets. They want to make sure that their personal information is collected, stored, processed, and shared by respecting their privacy preferences and data use agreements. Their main concern is that their genomic data could be misused in discriminatory ways or for individual identification in case of a data breach or privacy leak. This concern also includes their family members due to the inherent correlations between the genomes of family members. Therefore, for them, trust mainly depends on privacy and transparency.

Patients. Such individuals are the ones who benefit from the research outcomes (e.g., in terms of more personalized treatment). They want to make sure that they can benefit from the research results (e.g., considering the populations of data donors in the research dataset) and that the research is conducted in a correct way. Therefore, for them, trust mainly depends on representation and reproducibility.

Researchers who conduct the study. Such parties are the ones who create the research datasets, conduct the research, and disseminate the results. In the local setting (in Section IV.B.1), they want to make sure the shared research results do not result in additional privacy risks. In the collaborative setting (in Section IV.B.2), each collaborator (researcher) want to ensure the privacy requirements for their local dataset and they also want to make sure that collaborative research is done correctly (e.g., correctness of the computations that are done by other collaborators or a centralized server). Also, in both settings, researchers want to make sure that the data owners have trust in the system, so that they can construct large and diverse research databases. Therefore, for researchers who conduct the study, trust mainly depends on privacy, transparency, and reproducibility.

Research followers. Such parties are the ones who are only interested in the research findings (e.g., to use them in their own research initiatives). They want to make sure that the researchers who conduct the research share the results in a transparent way. They also want to make sure that the research is conducted correctly and that all populations can benefit from its findings. Therefore, for research followers, trust mainly depends on transparency, representation, and reproducibility.

V. CONCLUSIONS

In this paper, we highlighted requirements to support reproducibility, transparency, representation, and privacy in GWAS using technologies that provide privacy protection, quality control, and verifiability of the studies. We outlined conflicts across technical solutions and new research directions for different GWAS settings. This paper aimed thus at promoting technologies that would enhance the acceptance of genetic research. However, this paper does not discuss one of the greatest challenges of GWAS research: the

misappropriation or misinterpretation of GWAS results to promote or justify harmful or discriminatory ideologies. Some of the most impactful GWAS have involved studies of socially relevant traits and outcomes including educational attainment, intelligence, same-sex sexual behavior, and even household income. While there are no technical solutions to these challenges, some GWAS authors have begun to provide FAQs together with the study results. The purpose of such FAQs is precisely to advance socially responsible GWAS by explaining the nature of the study and implications of results in a manner accessible to broader audiences. "FAQs on Genomic Studies" (FoGS) is a repository of GWAS FAQs accessible to the public and designed to help mitigate misinterpretation and misapplication of socially sensitive GWAS research [53]. Enhancing trust and acceptance of GWAS results cannot be managed solely through technical solutions, and requires further study and interdisciplinary collaboration to develop comprehensive solutions. Finally, we are in a transition phase between execution of GWAS and a new generation of studies that use whole exome and genome sequencing data. As such, we view the experience in GWAS as important to the use of that technology, but also as a set of tools and principles that need to be implemented, and expanded to serve the unique features of sequencing data.

VI. ACKNOWLEDGEMENT

Research reported in this publication was supported by the National Institutes of Health under awards R35GM134927, R01LM014520, U54HG012510, R01LM013429, and R01LM013712, and by the National Science Foundation (NSF) under awards 2141622, 2050410, 2200255, and OAC-2112606. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research.

REFERENCES

[1] Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014 Jun;15(6):409–421. PMID:24805122

[2] Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet 2020 Jul;52(7):646–654. PMID:32601475

[3] Naveed M, Ayday E, Clayton EW, Fellay J, Gunter CA, Hubaux J-P, Malin BA, Wang X. Privacy in the Genomic Era. ACM Comput Surv [Internet] 2015 Sep;48(1). PMID:26640318

[4] Workflows [Internet]. [cited 2022 Apr 7]. Available from: https://www.myexperiment.org/workflows

[5] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004 Nov 22;20(17):3045–3054. PMID:15201187

[6] Bowers S, Ludäscher B. Actor-Oriented Design of Scientific Workflows. Conceptual Modeling – ER 2005 Springer Berlin Heidelberg; 2005. p. 369–384.

[7] Handigol N, Heller B, Jeyakumar V, Maziéres D, McKeown N. Where is the debugger for my software-defined network? Proceedings of the first workshop on Hot topics in software defined networks - HotSDN '12 [Internet] New York, New York, USA: ACM Press; 2012. [doi: 10.1145/2342441.2342453]

[8] Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013 Jan 18;339(6117):321–324. PMID:23329047

[9] Humbert M, Huguenin K, Hugonot J, Ayday E, Hubaux J-P. De-anonymizing genomic databases using phenotypic traits. Proceedings on Privacy Enhancing Technologies [Internet] 2015;2015(2). Available from: https://sciendo.com/downloadpdf/journals/popets/2015/2/article-p99.xml

[10] Humbert M, Ayday E, Hubaux J-P, Telenti A. Addressing the concerns of the lacks family: quantification of kin genomic privacy. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security New York, NY, USA: Association for Computing Machinery; 2013. p. 1141–1152.

[11] Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008 Aug 29;4(8):e1000167. PMID:18769715

[12] Humbert M, Ayday E, Hubaux J-P, Telenti A. On Non-cooperative Genomic Privacy. Financial Cryptography and Data Security Springer Berlin Heidelberg; 2015. p. 407–426.

[13] Atutornu J, Milne R, Costa A, Patch C, Middleton A. Towards equitable and trustworthy genomics research. EBioMedicine 2022 Feb;76:103879. PMID:35158310

[14] Mills MC, Rahal C. The GWAS Diversity Monitor tracks diversity by disease in real time. Nat Genet 2020 Mar;52(3):242–243. PMID:32139905

[15] Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc 2010 Sep;5(9):1564–1573. PMID:21085122

[16] Weale ME. Quality Control for Genome-Wide Association Studies [Internet]. Methods in Molecular Biology. 2010. p. 341–372. [doi: 10.1007/978-1-60327-367-1_19]

[17] Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G. Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray. Brief Funct Genomics 2016 Jul;15(4):298–304. PMID:26443613

[18] Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis [Internet]. International Journal of Methods in Psychiatric Research. 2018. p. e1608. [doi: 10.1002/mpr.1608]

[19] Duchi JC, Jordan MI, Wainwright MJ. Local Privacy and Statistical Minimax Rates. 2013 IEEE 54th Annual Symposium on Foundations of Computer Science 2013. p. 429–438.

[20] Kairouz P, Oh S, Viswanath P. Extremal Mechanisms for Local Differential Privacy [Internet]. arXiv [csIT]. 2014. Available from: http://arxiv.org/abs/1407.1338

[21] Dwork C. Differential Privacy: A Survey of Results. Theory and Applications of Models of Computation Springer Berlin Heidelberg; 2008. p. 1–19.

[22] Erlingsson Ú, Feldman V, Mironov I, Raghunathan A, Talwar K, Thakurta A. Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity [Internet]. Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms. 2019. p. 2468–2479. [doi: 10.1137/1.9781611975482.151]

[23] Cheu A, Smith A, Ullman J, Zeber D. Distributed differential privacy via shuffling. Conference on the Theory … [Internet] Springer; 2019; Available from: https://link.springer.com/chapter/10.1007/978-3-030-17653-2_13

[24] Murakami T, Kawamoto Y. ${Utility-Optimized}$ Local Differential Privacy Mechanisms for Distribution Estimation. 28th USENIX Security Symposium (USENIX Security 19) 2019. p. 1877–1894.

[25] Gu X, Li M, Xiong L, Cao Y. Providing Input-Discriminative Protection for Local Differential Privacy [Internet]. arXiv [csCR]. 2019. Available from: http://arxiv.org/abs/1911.01402

[26] Yilmaz E, Ji T, Ayday E, Li P. Genomic Data Sharing under Dependent Local Differential Privacy [Internet]. arXiv [csCR]. 2021. Available from: http://arxiv.org/abs/2102.07357

[27] Humbert M, Ayday E, Hubaux J-P, Telenti A. Reconciling Utility with Privacy in Genomics. Proceedings of the 13th Workshop on Privacy in the Electronic Society New York, NY, USA: Association for Computing Machinery; 2014. p. 11–20.

[28] Yilmaz E, Ji T, Ayday E, Li P. Preserving Genomic Privacy via Selective Sharing. Proc ACM Workshop Priv Electron Soc 2020 Nov;2020:163–179. PMID:34485998

[29] Cho H, Wu DJ, Berger B. Secure genome-wide association analysis using multiparty computation. Nat Biotechnol Nature Publishing Group; 2018 May 7;36(6):547–551.

[30] Huang Z, Lin H, Fellay J, Kutalik Z, Hubaux J-P. SQC: secure quality control for meta-analysis of genome-wide association studies. Bioinformatics 2017 Aug 1;33(15):2273–2280. PMID:28379351

[31] Lu W-J, Yamada Y, Sakuma J. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. BMC Med Inform Decis Mak 2015 Dec 21;15 Suppl 5:S1. PMID:26732892

[32] Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, Berger B, Fellay J, Hubaux J-P. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nat Commun 2021 Oct 11;12(1):5910. PMID:34635645

[33] Kim D, Son Y, Kim D, Kim A, Hong S, Cheon JH. Privacy-preserving approximate GWAS computation based on homomorphic encryption. BMC Med Genomics 2020 Jul 21;13(Suppl 7):77. PMID:32693801

[34] Constable SD, Tang Y, Wang S, Jiang X, Chapin S. Privacy-preserving GWAS analysis on federated genomic datasets. BMC Med Inform Decis Mak 2015 Dec 21;15 Suppl 5:S2. PMID:26733045

[35] Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, Weiss S, Völker U, Pitkänen E, Heider D, Wenke NK, Kaissis G, Rueckert D, Kacprowski T, Baumbach J. sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. Genome Biol 2022 Jan 24;23(1):32. PMID:35073941

[36] Kockan C, Zhu K, Dokmai N, Karpov N, Kulekci MO, Woodruff DP, Sahinalp SC. Sketching algorithms for genomic data analysis and querying in a secure enclave. Nat Methods 2020 Mar;17(3):295–301. PMID:32132732

[37] NOT-OD-19-023: Update to NIH Management of Genomic Summary Results Access [Internet]. [cited 2022 Apr 8]. Available from: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html

[38] Fienberg SE, Slavkovic A, Uhler C. Privacy Preserving GWAS Data Sharing [Internet]. 2011 IEEE 11th International Conference on Data Mining Workshops. 2011. [doi: 10.1109/icdmw.2011.140]

[39] Yu F, Fienberg SE, Slavković AB, Uhler C. Scalable privacy-preserving data sharing methodology for genome-wide association studies [Internet]. Journal of Biomedical Informatics. 2014. p. 133–141. [doi: 10.1016/j.jbi.2014.01.008]

[40] Johnson A, Shmatikov V. Privacy-Preserving Data Exploration in Genome-Wide Association Studies. KDD 2013 Aug;2013:1079–1087. PMID:26691928

[41] Tramèr F, Huang Z, Hubaux J-P, Ayday E. Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security New York, NY, USA: Association for Computing Machinery; 2015. p. 1286–1297.

[42] Shringarpure SS, Bustamante CD. Privacy Risks from Genomic Data-Sharing Beacons. Am J Hum Genet 2015 Nov 5;97(5):631–646. PMID:26522470

[43] von Thenen N, Ayday E, Cicek AE. Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics 2019 Feb 1;35(3):365–371. PMID:30052749

[44] Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods 2016 Mar;13(3):251–256. PMID:26828419

[45] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet 2011 Jan;Chapter 1:Unit1.19. PMID:21234875

[46] Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, Fullerton SM, Garrison NA, Ghosh N, Hammonds EM, Jones DS, Kenny EE, Kraft P, Lee SS-J, Mauro M, Novembre J, Panofsky A, Sohail M, Neale BM, Allen DS. Getting Genetic Ancestry Right for Science and Society [Internet]. arXiv [q-bioPE]. 2021. Available from: http://arxiv.org/abs/2110.05987

[47] Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, Ferreira T, Fall T, Graff M, Justice AE, Luan J 'an, Gustafsson S, Randall JC, Vedantam S, Workalemahu T, Kilpeläinen TO, Scherag A, Esko T, Kutalik Z, Heid IM, Loos RJF, Genetic Investigation of Anthropometric Traits (GIANT) Consortium. Quality control and conduct of genome-wide association meta-analyses. Nat Protoc 2014 May;9(5):1192–1212. PMID:24762786

[48] Walfish M, Blumberg AJ. Verifying computations without reexecuting them. Commun ACM New York, NY, USA: Association for Computing Machinery; 2015 Jan 28;58(2):74–84.

[49] Yu X, Yan Z, Vasilakos AV. A survey of verifiable computation. Mob Netw Appl Springer Science and Business Media LLC; 2017 Jun;22(3):438–453.

[50] Gennaro R, Wichs D. Fully Homomorphic Message Authenticators. Advances in Cryptology - ASIACRYPT 2013 Springer Berlin Heidelberg; 2013. p. 301–320.

[51] Dervishi L, Wang X, Li W, Halimi A, Vaidya J, Jiang X, Ayday E. Facilitating Federated Genomic Data Analysis by Identifying Record Correlations while Ensuring Privacy [Internet]. arXiv [csCR]. 2022. Available from: http://arxiv.org/abs/2203.05664

[52] Halimi A, Dervishi L, Ayday E, Pyrgelis A, Troncoso-Pastoriza JR, Hubaux J-P, Jiang X, Vaidya J. Privacy-Preserving and Efficient Verification of the Outcome in Genome-Wide Association Studies [Internet]. arXiv [csCR]. 2021. Available from: http://arxiv.org/abs/2101.08879

[53] Martschenko DO, Domingue BW, Matthews LJ, Trejo S. FoGS provides a public FAQ repository for social and behavioral genomic discoveries. Nat Genet 2021 Sep;53(9):1272–1274. PMID:34493865