GenoShare: Supporting Privacy-Informed Decisions for Sharing Exact Genomic Data

Jean Louis Raisaro
EPFL

Carmela Troncoso
IMDEA Software Institute

Mathias Humbert
Swiss Data Science Center

Zoltan Kutalik
Lausanne University Hospital

Amalio Telenti
Human Longevity Inc.

Jean-Pierre Hubaux
EPFL

ABSTRACT

The academic community has proposed many solutions to address the privacy concerns associated with genomic-data sharing. However, practitioners have not adopted these solutions due to their impact on the data utility. To address this problem, we introduce GenoShare, a framework that helps practitioners to make informed decisions about the sharing of *exact genomic data* by providing means to systematically reason about the risk of disclosing privacy-sensitive attributes (e.g., health status, kinship, physical traits). We instantiate GenoShare with three of the most important genomics-oriented inference attacks, and demonstrate its capability to detect potential leakage of sensitive attributes using real data from the 1000 Genomes Project.

1. INTRODUCTION

The privacy risks stemming from disclosing medical genomic data [13, 17, 40] are being increasingly amplified by the growing number of breaches occurring in healthcare organizations [53, 44, 21, 56]. This situation creates a complicated environment for health care practitioners and researchers trying to engage with citizens regarding the sharing of data for clinical research, as gaining their trust is becoming a major challenge. Currently, medical institutions and research centers address this problem by relying on a review board that decides whether disclosure is suitable. However, these decisions usually follow all-or-nothing policies, which provide little control on the inferences that can be made upon the shared data. Thus, they are of little help at conveying trust to users. The computer security community has made a remarkable effort to improve this situation, mainly focusing on solutions that perturb the data such that releases are differentially private [15, 28, 62], since anonymization approaches [51, 35, 33, 61] have been shown futile for privacy-preserving sharing of genomic data [18, 22].

Despite the demanding privacy needs of genomic data management, these solutions have not been adopted by practitioners so far. A main reason for this reluctance is that genomics applications usually require genomic data to be as exact as possible [13, 16, 41]. High accuracy is especially important in association studies aiming to identify significant correlations between particular genotypes and rare diseases, which are often weak signals. Moreover, differentially private solutions focus on safeguarding only the release of aggregates, and thus are not suitable for protecting individual's data, whose sharing is a common practice in research studies. In summary, *the need to release the* ***exact data values*** *precludes the use of state-of-the-art solutions that provide formal privacy guarantees in the presence of arbitrary external information.*

Yet, genomic data sharing is crucial to advance the state of the art in medicine. Thus, there is a high demand in the biomedical community for solutions that enable practitioners to reason about what exact data can be released while protecting individuals' privacy in clinical and research settings. Even though they cannot prevent inferences enabled by unforeseeable attack developments or data releases, such solutions would represent a great improvement over the current situation since they can effectively reduce the privacy risks based on the information available to the decision maker.

In this paper, we introduce GenoShare, whose goal is to assist practitioners in decision making by quantifying the risk of sensitive information leakage when sharing genomic data. Let us assume that an institution (e.g., hospital, research center) wishes to share genomic data, but is concerned about the privacy of the individuals who contributed their data. Upon reception of a request for genomic data sharing, such institution can use GenoShare to quantify the risk of sensitive attribute disclosure associated to revealing those data. To this end, GenoShare considers inference attacks relevant to the privacy concerns of the data contributors, and the information available to the adversary: i) the genomic statistics across populations [52], ii) the genomic association to sensitive information [20, 50], and iii) the correlations between genomic variants inside an individual's genome, and across related individuals' genomes. As opposed to prior works that consider only one type of inference attack at a time [22, 25, 46], GenoShare quantifies the risk of a privacy breach considering the joint effect of inference attacks, i.e., exploiting their interrelations, and can also consider partial adversarial knowledge – thus providing a more realistic risk estimation than the state-of-the-art

approaches.

If the risk of sensitive attribute disclosure is deemed low, the institution can release the requested data in exact form, and otherwise it denies access to the data. GenoShare measures risk using novel intuitive metrics that, as opposed to current approaches based on inferences of raw genomic values [58], are directly related to the inference of tangible information, such as kinship or predisposition to a disease. Thus, they are well suited for modeling informed consent [31]. Furthermore, since denying access based on information secret to the adversary is known to leak information [29], GenoShare implements mechanisms to avoid this leakage.

To summarize, we make the following contributions:

✓ We present GenoShare, a framework that supports informed decision making regarding the sharing of *exact* genomic data by considering relevant inference attacks, and their joint effect on privacy.

✓ We introduce novel metrics that capture the risk of sensitive attributes disclosure, better suited to model informed consent than the state of the art.

✓ We develop a novel method for preventing inferences based on genomic query denials. The idea is to internally use *avatars* (modified versions of individual's genomes) to decide upon data release, still releasing the original data when privacy is not at risk.

✓ We instantiate GenoShare with the three most relevant attacks on genomic privacy, advancing the state of the art by adapting them to consider partial information and considering their interrelations to amplify their inference power. We show GenoShare's effectiveness at detecting potential private information leaks using real data from the 1,000 Genomes Project [52].

2. GENOMICS 101

Genetic Variation. The human genome consists of three billion pairs of nucleotides with values in the set $\{A, T, C, G\}$. Around 99.5% of the whole human genome is identical for any two individuals, and the remaining part is referred to as *genetic variation*. Out of the many existing types of genetic variations, we focus on the most common, which stems from differences in single nucleotides, called *single nucleotide variants* (SNVs). In the human population, a given genetic locus (defined as a position in a chromosome) can have several possible versions (or alleles). Each individual either has two copies of the same allele (*homozygous*) or two copies of different alleles (*heterozygous*). Genetic variants in a given individual genome are identified by comparing the genome with the *reference human genome*, a digital sequence of nucleotides considered representative of the human genetic makeup. In the vast majority of cases, a genetic variant is biallelic, i.e., it can take two different alleles: a reference allele, the one appearing on the reference human genome, and an alternate allele, the alternative version occurring in the human population. The latter presence is quantified by the *alternate allele frequency*, or aaf. Hence, a genetic variant, at a given position, can be homozygous reference (i.e., taking two reference alleles), heterozygous (one reference and one alternate alleles) or homozygous alternate (two alternate alleles). We encode the value (or genotype) of a variant g_i at position i as $g_i \in \{0, 1, 2\}$, based on the number of alternate alleles it contains.

Genotype-Phenotype Association. Genomic variants can be associated to phenotypes, e.g., diseases or physical traits, either increasing the predisposition of an individual to develop a disease at some point in time, or being protective with respect to that disease. The strength of this association is generally quantified by the *effect size*, denoted as $\omega = log(OR)$, where OR is the *odds ratio*. The *odds* represent the ratio between the probability of disease occurrence in a given group and the probability of non-occurence in the same group. The *odds ratio* is the odds in the group of individuals carrying a genetic variant divided by the odds in the group of those not carrying it. If there are N_{dg} individuals carrying a disease and a variant, N_{hg} healthy individuals carrying the same variant, N_{dn} individuals carrying the disease but not the variant, and N_{hn} healthy individuals not carrying the variant, then the OR is $\frac{N_{dg}/N_{hg}}{N_{dn}/N_{hn}}$.

Genetic Correlations. Because genetic segments (or haplotypes) are inherited in blocks, physically close variants are very often correlated. Such a correlation is called *linkage disequilibrium* (LD). Beyond intra-genome correlations, there exist inter-genome correlations that stem from reproduction. During the reproduction process, at each genetic position, a child inherits one allele from his mother and one from his father. Under the Mendelian inheritance assumption, each allele of a parent is passed to the child with equal probability 0.5, independently of the other positions. Moreover, given both parents' genomes, a child's genome is conditionally independent of all other ancestors' genomes.

3. THE GENOMIC SHARING SCENARIO

We consider a scenario in which an institution holds a database $\mathbf{D}$ with genomic data of lots of individuals. We model an individual's genome in $\mathbf{D}$ as a vector $\mathbf{g} = (g_1, \ldots, g_n)$ formed by n variants on autosomal chromosomes (i.e., not sex chromosomes), where g_i denotes the value of variant i. We use a vector $\mathbf{f} = (f_1, \ldots, f_n)$ to model aggregated statistics on these variants. We summarize the notation used throughout the paper in Table 2 (Appendix D).

Institutions wish to share these data in one of two ways: i) as a subset of genotypes $\mathbf{g}_s$, in response to a *genotype request* for variants of a given individual, $q_g(\mathbf{g}_s)$; and ii) as a subset of aggregated statistics $\mathbf{f}_s$ for a specific group of individuals, in response to an *aggregated request* for variants, $q_m(\mathbf{f}_s)$. Because genomic-related applications are not tolerant to noisy data, institutions want to share them in their original, exact, form.

On the other hand, individuals whose genomes are in $\mathbf{D}$ could be concerned about the potential disclosure of their sensitive information. They express such concerns establishing a threshold, $\boldsymbol{\rho}$, that captures their tolerance to disclosure of sensitive information with respect to the risk of inference.

We assume an adversary who wants to learn some (sensitive) information about individuals in the database protected by GenoShare. Genome-based inference attacks can be categorized as follows: (i) *Phenotype inference attacks*, that aim at inferring an individual's predisposition to diseases (e.g., Alzheimer's disease, cancer, schizophrenia) [13], or her physical traits, from her genotype and known genotype-phenotype correlations [20]; (ii) *Membership inference attacks*, whose goal is to infer the presence of an individual of whom genomic information is available in a group for which aggregate statistics are known [22, 48], which can be

very sensitive if such a group is associated with a sensitive attribute (e.g., HIV-positive patients, patients in a psychiatric institute, etc.); (iii) *Kinship inference attacks*, that aim at inferring familial relationships between know individuals' genomes; (iv) *Re-identification attacks*, that aim at inferring the identity (e.g., family name) [18] behind a known genome, or physical traits (e.g., height, eye color, etc.) that can lead to re-identification [9]; or (v) *Linking attacks*, that aim at linking anonymized sensitive phenotype data available to the adversary to a set of individuals for which their genotypes are known by exploiting genotype-phenotype correlations [26, 19].

To perform inferences, the adversary could have access to the following information :
–Background information ($\mathcal{B}$) such as public information about average individuals' genomes [52], and about genomic association to sensitive information [50, 20]; or information made public by individuals, e.g., on OpenSNP [42] that provides access to further genomic data, or on Facebook [14] that provides information about familial relationships [25],
–Revealed variants of the targeted individual and of her relatives ($\mathbf{A_g}$), and aggregated statistics ($\mathbf{A_m}$),
–Potentially revealed variants: information that would be revealed if a new request is granted ($\mathbf{g}_s$ or $\mathbf{f}_s$), or that could be inferred in case of GenoShare denying a high-risk query.

4. GENOSHARE

We design GenoShare to help institutions owning a database of genomic sequences to share exact genotypes ($\mathbf{g}_s$) or aggregated statistics ($\mathbf{f}_s$) in a privacy-conscious way. When GenoShare receives a data request (either $q_g(\mathbf{g}_s)$ or $q_m(\mathbf{f}_s)$), it quantifies the privacy risks regarding inferences stemming from the release of the data. If the risks are deemed too high with respect to given thresholds $\boldsymbol{\rho}$, GenoShare prevents any automatic release of data. It can further provide the institution with information that can be used to make a privacy-conscious decision regarding whether to share the requested data.

We note that when queries are granted, GenoShare *cannot protect the information that has already been released* from inferences that could be made possible by advances in the state of the art in genomics research. This limitation is inherent to the practitioners' need for clean and exact data.

4.1 Architecture

GenoShare is conceptually divided in four main blocks, illustrated in Figure 1:
Attack Engine. This engine simulates the inference attacks that the adversary can perform given both the information already available to him ($\mathbf{A_g}$, $\mathbf{A_m}$, and $\mathcal{B}$), and what would be disclosed if the query was granted ($\mathbf{g}_s$ or $\mathbf{f}_s$).
Risk Measurement Engine. This engine computes the risk of sensitive attribute disclosure materializing if the data requested in the query is revealed.
Decision Engine. This engine checks if the risk computed by the Risk Measurement Engine exceeds the established thresholds for any individual in the database.
Avatar Engine. The Avatar Engine creates and stores avatars: modified versions of the stored genomic data. Avatars are used, *internally*, in the Attack Engine to simulate the inference attacks, and in the Risk Measurement Engine to quantify the inference risk. Their goal is to mitigate potential inferences on the true genome based on query denials. Note that, whenever the decision is to grant the queried data, the *true* data is released.

4.2 Using GenoShare

Initialization. GenoShare requires an initialization phase in which the attacks to be considered are instantiated in the Attack Engine (Sect. 5); the privacy metrics are set up in the Risk Measurement Engine (Sect. 6); the corresponding risk thresholds $\boldsymbol{\rho}$ are set up in the Decision Engine; and the Avatar Engine generates one avatar per individual per attack considered in the Attack Engine (Sect. 7).
Operation. Upon reception of a data request ($q_g(\mathbf{g}_s)$, or $q_m(\mathbf{f}_s)$), two steps are needed to run GenoShare:
1. Configuration. GenoShare needs to be configured to decide: (i) what background information $\mathcal{B}$ is assumed to be available to the adversary (e.g., only data released by tool, familial relationships, genome data obtained from other sources,...); and (ii) which attacks are of a concern for the given request.
2. Execution. To evaluate the query, GenoShare substitutes the requested data ($\mathbf{g}_s$ or $\mathbf{f}_s$) by the corresponding avatar genotypes ($\tilde{\mathbf{g}}_s$ or $\tilde{\mathbf{f}}_s$). Then, it runs *all* the attacks configured in the Attack Engine, whose output is then input to the Risk Measurement Engine. This engine computes the risk of a privacy breach, and the Decision Engine verifies whether it complies with all individuals' thresholds. If *any* of the risks exceeds the corresponding thresholds, GenoShare prevents the release of the data. If not, it releases the exact data requested in the query ($\mathbf{g}_s$ or $\mathbf{f}_s$).

5. GENOSHARE'S ATTACK ENGINE

The Attack Engine runs all the inference attacks configured in GenoShare , to mimic the actions that an adversary would perform to learn sensitive information about individuals. As opposed to prior works that consider inference attacks independently, GenoShare considers them jointly, thus maximizing their effect. We now introduce the three most well-known attacks, namely *phenotype*, *membership*, and *kinship inference*, that we instantiate with state-of-the-art inference techniques. We note that GenoShare can accommodate other attacks, such as the *re-identification attack*, the *linking attack* or others, and it can be updated with better techniques each time there is a new proposal.

5.1 Phenotype inference attack

Phenotype inference attacks are aimed at learning genotype-related sensitive phenotypes about a target individual (and her relatives), such as predisposition to diseases or physical traits. For simplicity, in this paper we only consider predisposition to diseases inference.

Phenotype inference attacks run in two steps: (i) *genotype completion*, in which the adversary uses the target's known variants, i.e., those that he has already observed, to infer correlated unobserved variants both from the target and her relatives, and (ii) *genotype-phenotype mapping*, in which, given the recovered variants, the adversary computes the target's disease predisposition using publicly available information about genotype-phenotype correlations. We now provide details about these two phases that are relevant to understand the avatar generation algorithms in Sect. 7.
Genotype completion. Genotype completion enables the inference of unobserved variants $\mathbf{g}_u$ from the variants available to the adversary (i.e., previously revealed and to be

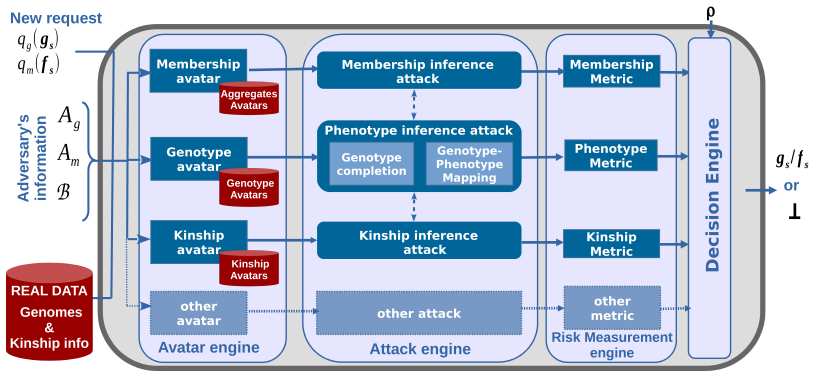


Figure 1: The four main blocks in GenoShare: i) an *Avatar Engine* that generates and stores avatar genomes which are used *internally* to avoid inferences based on query denials; ii) an *Attack Engine* that simulates the adversary's behavior in order to predict what information could be learned if the requested data are released; iii) a *Risk Measurement Engine* that quantifies the risk of disclosing sensitive information when releasing data; and iv) a *Decision Engine* that verifies whether the privacy risks are under the thresholds $\boldsymbol{\rho}$, and either outputs the **exact requested data** ($\mathbf{g}_s$ or $\mathbf{f}_s$) or nothing. The figure shows instantiations of the blocks for three inference attacks, and the bottom row illustrates how other attacks could be accommodated.

revealed, $\mathbf{g}_o := \mathbf{A}_\mathbf{g} \cup \mathbf{g}_s$, and background information $\mathcal{B}$), and the genotype data in a public panel of reference individuals $\mathbf{R}$. It outputs a posterior probability distribution for each unobserved variant, $\hat{\mathbf{g}} = \Pr(\mathbf{g}_u|\mathbf{g}_o, \mathcal{B}, \mathbf{R})$. We instantiate this inference using a well-established statistical technique called *genotype imputation* [37] that makes use of a Hidden Markov Model (HMM) to model the target's genome. Simpler techniques could be used but, to the best of our knowledge, genotype imputation provides the most accurate genotype inference [46].

Then, the adversary can infer the variants of the target's relatives from the completed genotype, i.e., the exact observed genotypes $\mathbf{g}_o$ and the probabilistically inferred ones $\hat{\mathbf{g}}$, and the knowledge on target's family tree. We model the target's family as a Bayesian network and the well-established *junction tree* algorithm [32] can be used to compute, for each family member, the set of marginal probability distributions over the unobserved variants conditioned on the target's information previously inferred ($\mathbf{g}_o \cup \hat{\mathbf{g}}$).

The details of both genome imputation and the junction tree algorithm are explained in Appendix A.

Genotype-phenotype mapping. An individual's genetic predisposition to a disease can be inferred from her genotype at variants associated with the disease, and the strength of this genotype-phenotype association. As explained in Section 2, this strength is characterized by the effect size $\omega^y = \log(OR)$, where OR is the odds ratio. More formally, let $\Psi(y)$ be the set of variants associated with disease y. Then, the adversary computes the target's predisposition to disease y, denoted as P^y, as the linear combination of the target's inferred genotypes $\hat{g}_i$ of variants i in $\Psi(y)$ weighted by the strength of the genotype-phenotype association ω_i.

5.2 Membership inference attack

The membership inference attack enables the adversary to infer whether a target individual, for which variants are known, is present in a group of individuals for which genetic aggregated statistics are available [22, 59, 47, 57, 27, 64, 48]. We instantiate this attack using the technique proposed by Homer et al. [22] because it relies on less restrictive assumptions than other approaches, but any other membership inference technique could be used instead. This technique compares, for every target's observed variants $\mathbf{g}_o$, (i) the distance between the alternate allele frequency $\frac{g_i}{2}$ in the target's genotype and aaf_i (the alternate allele frequency in the population) with (ii) the distance between $\frac{g_i}{2}$ and f_i, the frequency of the same allele in the group of interest. Formally, using the L_1 distance:

$$D\left(\frac{g_i}{2}\right) = \left\|\mathsf{aaf}_i - \frac{g_i}{2}\right\| - \left\|f_i - \frac{g_i}{2}\right\|. \tag{1}$$

When the target is in the group, $E\left[D(\frac{g_i}{2})\right]$ is greater than zero because $\frac{g_i}{2}$ shifts f_i away from aaf_i. On the contrary, under the null hypothesis (the target is not present in the group of interest) $E\left[D(\frac{g_i}{2})\right]$ should approach zero. If $\frac{g_i}{2}$ is further away from the group than from the reference population, i.e., even less likely to be part of the group, $E\left[D(\frac{g_i}{2})\right]$ is negative.

If the number of released frequencies is sufficiently high, $E\left[D(\frac{g_i}{2})\right]$ converges to the normal distribution due to the central limit theorem. This enables the adversary to make use of a one-sample t-test to determine whether the target is part of the group or not. As we explain in Section 5.4, the adversary can make use of the output of the *genotype completion* ($\hat{\mathbf{g}}$) to further improve the membership attack. As the output of genotype completion is a probability distribution over the possible values of a variant, we adapt (1) such that it incorporates this knowledge. For $k \in \{0, 1, 2\}$:

$$D\left(\frac{\hat{g}_i}{2}\right) = \left\|\mathsf{aaf}_i - \sum_k \frac{k}{2}\Pr(\hat{g}_i = k)\right\| - \left\|f_i - \sum_k \frac{k}{2}\Pr(\hat{g}_i = k)\right\|. \tag{2}$$

5.3 Kinship inference attack

The kinship inference attack enables the adversary to infer the degree of kinship of a pair of target individuals, given a common set of their variants. We instantiate this attack, for the first time, with a technique that estimates the proportion of genomic variants co-inherited from a common ancestor [36, 54]. Similarly to the previous attacks, other

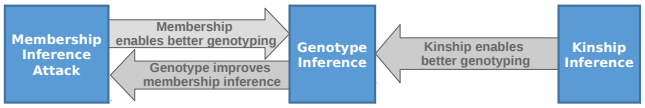


Figure 2: Interrelation among inference techniques

techniques could be used. In particular, we use the kinship coefficient $\phi_{A,B}$ proposed in [36], defined as the probability that two alleles sampled at random from two individuals A and B are identical by descent. We compute $\phi_{A,B}$ as the average over all variants' coefficients $\hat{\phi}_{i,A,B}$ computed as:

$$\hat{\phi}_{i,A,B} = \frac{2\mathsf{aaf}_i(1-\mathsf{aaf}_i) - (g_{i,A} - g_{i,B})^2}{8\sum_{i \in M_{A,B}} \mathsf{aaf}_i(1-\mathsf{aaf}_i)}, \tag{3}$$

where $M_{A,B}$ is the set of observed variants for both individuals. If the number of observed variants is sufficiently high, the sum of $\hat{\phi}_{i,A,B}$ over all variants converges to the normal distribution due to the central limit theorem. Similarly to the previous attack, the adversary can infer the degree of kinship of the target individuals by using the inference criteria in [36] (see Table 1 in Appendix B). As there are different kinship levels, in our experiments we use a closed testing procedure [38] of consecutive one-sample t-tests and choose the closest relationship that can be inferred.

5.4 Attacks interrelations

In order to best estimate the adversary's inference capabilities, the Attack Engine takes into account interrelations between the different attacks which can benefit from each other as shown in Figure 2. The genotype completion carried out within the phenotype inference attack can be used to improve the efficacy of the membership inference attack, because it increases the knowledge of the adversary about the target's genotype information: A larger number of the target's variants is made available to establish her membership to the database. On the contrary, the kinship inference attack cannot benefit from genotype completion. This is because genotype completion relies on correlations between variants, but the accuracy of the estimated kinship coefficient relies on independent variants. However, the kinship inference attack can improve the phenotype inference attack by informing about the familial ties which enable us to build the Bayesian network model. As a consequence, the kinship inference attack indirectly and positively influences the membership inference attack. Finally, the membership inference attack can also enhance the genotype completion of the phenotype inference attack if it reveals that the target is present in a database associated with a phenotype that is correlated to a particular genotype. We do not evaluate the latter in Section 8, as understanding the information gained by the adversary is straightforward.

6. GENOSHARE'S RISK MEASUREMENT ENGINE

GenoShare needs to have means to measure the risk of sensitive attribute disclosure when granting a request $q_g(\mathbf{g}_s)$ or $q_m(\mathbf{f}_s)$. Such risk needs to be understood by a large variety of users with extremely diverse knowledge related to genomics and/or medicine (e.g., patients, doctors, researchers). As such, we propose metrics to represent sensitive attributes that could be understood by the public at large [31]. It must be noted, however, that the Risk Measurement Engine could be instantiated with any other metrics deemed suitable for measuring the information leaked to the adversary in order to characterize other more-or-less-specialized concerns. For instance, metrics oriented to avoid bulk disclosure of data, e.g., by including a large percentage of variants in the risk computation, or metrics of interest for experts, such as genomic researchers, in which only specific variants are considered to be risky.

6.1 Phenotype inference risk

The phenotype inference risk aims at capturing how well the adversary can infer a target's phenotype, e.g., a physical trait or a predisposition P^y to a disease, regardless of the actual phenotype value, i.e., her inference error. For simplicity we focus our explanation on disease predisposition, but note that the metric can be easily adapted to other phenotype inferences.

Disease predisposition can be inferred through the *phenotype inference attack*, see Section 5.1. If the adversary had access to all the target's actual genotypes, he could perfectly compute her predisposition. However, when GenoShare is in place, the adversary only has access to the previously disclosed genotypes ($\mathbf{A_g}$), and to those that he can infer using the genotype completion ($\hat{\mathbf{g}}$). Recall from Section 5.1 that, for each $\hat{g}_i$, the adversary obtains a probability distribution over the three possible variant values $\{0,1,2\}$. Given the known and inferred variants, the predisposition P^y can be estimated using the genotype-phenotype mapping.

Wagner defined the per-variant success rate of the adversary as the probability that the adversary correctly infers the true variant value given the genotype inference output [58]. Inspired by this metric, we quantify the risk of inferring a given disease predisposition, denoted as R^y, by weighing the per-variant privacy metric by the per-variant strength of association ω_i between genotype and disease predisposition:

$$R^y = \frac{1}{\sum_i \omega_i} \sum_i \omega_i \Pr(\hat{g}_i = g_i), \quad i \in \Psi(y), \tag{4}$$

where $\Psi(y)$ is the set of variants associated with disease y and g_i the individual's true genotype of variants i in $\Psi(y)$.

6.2 Membership and kinship inference risks

We consider membership and kinship inferences to be binary classification problems. These are based on a one-sample t-test used to test the null hypothesis of the individual not being in the dataset, in the case of membership (see Sect. 5.2), or not being related to anyone else in the database in the case of kinship (see Sect. 5.3). Thus, to quantify these inference risks, we use α, the significance level of the classification (i.e., the false positive rate), and $1-\beta$, the test statistical power, where β denotes the false negative rate. Intuitively, the higher the power and the lower the significance level are, the more certain the adversary is about his classification. Therefore, an individual's privacy grows with α and β (i.e., when the number of false positives, resp. false negatives, grows).

We define the risk of membership inference as

$$R^m = (1-\beta_m, \alpha_m)\,, \tag{5}$$

and the risk of inferring kinship of degree d as

$$R^k_d = (1-\beta_d, \alpha_d), \quad d \in \mathbb{N}^*\,. \tag{6}$$

7. GENOSHARE'S AVATAR ENGINE

Denying access to private data can leak information about these data, because decisions are based on information not available to the attacker [29]. Simulatable auditing [30] prevents such leakage by anticipating incoming queries, and only replying to those that do not enable unauthorized inferences. Unfortunately, existing solutions in this direction are limited to statistical queries different from aggregated requests in the genomic scenario, and not applicable to the case of individual genotype requests.

The key intuition in simulatable auditing is that, to prevent leakage, a decision to deny a query must be based exclusively on the information released by the system (including the potential answer to the current query), but *not* on the query itself nor other value in the database. Building on this idea, one may be tempted to use existing techniques to produce synthetic data [5], or perturb the data to make it differentially private [2], in order to obtain alternative data with similar statistical properties to those of the original data. These alternative data can be used as input for the decision process so that the query denial does not depend on the original sensitive data. However, by following this approach, GenoShare can take incautious decisions in particular instances, i.e., granting a query deemed safe for the alternative input, while for the original genomic data it would have raised an alarm. Such a risky behavior is not acceptable for medical institutions notably because of patients' privacy.

To mitigate this problem, we propose to use *avatars*, new modified versions of an individual's genome or a database's aggregates used *internally* by GenoShare. Avatars, as opposed to perturbed or synthetic data, always guarantee conservative decisions when used in GenoShare's decision process. They are used as input to the Attack and Risk Measurement Engines instead of the original genomes, thus ensuring that, given a denial, the adversary can, at most, recover the avatar. We note that when a query is granted, i.e., deemed safe, always the original data is released, not the avatars.

We define two types of avatars: genome avatars ($\tilde{\mathbf{g}}$), and aggregates avatars ($\tilde{\mathbf{f}}$) to substitute genotypes ($\mathbf{g}_s$) and aggregates ($\mathbf{f}_s$) real inputs, and construct them to ensure that decisions are never incautious. In terms of the risk metrics defined in Section 6, this implies that *phenotype inference attacks* on the avatar should result in a success rate, R^y, larger than on the real genome; and *membership* (resp. *kinship*) *inference attacks* should result in higher power, $1-\beta_m$ (resp. $1-\beta_d$) for a given significance level α_m (resp. α_d).

Avatar-based privacy. Guaranteeing safety of the decisions with respect to the real genomes inevitably leads us to generating avatars that are not fully independent from the real genomic data they represent. Thus, we cannot provide the provable protection guaranteed by simulatable auditing. Instead, we quantify the level of privacy provided by GenoShare's avatars. Let us consider genotype requests as an example. Given a query denial, the probability of the adversary inferring one true genotype is:

$$\Pr[g_i|\tilde{g}_i] \cdot \Pr[\tilde{g}_i|\text{denial}]\,, \tag{7}$$

where $\tilde{g}_i$ denotes the value of g_i's avatar, $\Pr[g_i|\tilde{g}_i]$ denotes the probability that the adversary succeeds at recovering true genotypes from the avatar, and $\Pr[\tilde{g}_i|\text{denial}]$ denotes the probability of learning the avatar from the denial. The latter strongly depends on the concrete sequence of queries and their replies, hence cannot be computed analytically. Thus, we choose to assume the worst-case scenario in which the adversary does recover the avatar and concentrates on computing the first probability, $\Pr[g_i|\tilde{g}_i]$. This worst-case scenario provides a lower bound on the privacy provided by avatars. If the adversary cannot correctly recover the avatar from the denial (i.e., $\Pr[\tilde{g}_i|\text{denial}] < 1$), the overall privacy increases.

Then, for a given individual, we compute her avatar's privacy as the average error of the adversary over all variants:

$$\mathsf{Priv}_{\tilde{\mathbf{g}}} = \frac{1}{n}\sum_i \left(1 - \Pr[g_i|\tilde{g}_i]\right), \quad \mathsf{Priv}_{\tilde{\mathbf{f}}} = \frac{1}{n}\sum_i \left(1 - \Pr[f_i|\tilde{f}_i]\right). \tag{8}$$

We note that, depending on the use case, it could make sense to only consider variants that are deemed most sensitive for the individual.

In the following, we propose avatar-generation algorithms for the three families of techniques we instantiate in the Attack Engine. We note that the proposed avatar generation methods are not tied to any particular implementation of the attacks, but based on their fundamental operation principle. Thus, they are valid for any attack inside a family.

7.1 Genome avatar

Genome avatars $\tilde{\mathbf{g}}$ are used as input to the Attack Engine when GenoShare receives a genomic request $q_g(\mathbf{g}_s)$. Since the inference techniques are based on different principles, avatars must be technique-dependent to guarantee that, for all cases, GenoShare outputs a conservative decision.

Phenotype inference. GenoShare quantifies the phenotype inference risk stemming from a phenotype inference attack using R^y (as in (4)), dependent on the adversary's error. Hence, to trigger conservative decisions avatars need to reduce this error with respect to the case where the real genome would be used for the attacks. Phenotype inference attacks rely on genome completion to infer unknown variants before using a phenotype-genotype mapping to perform the inference (see Sect. 5.1). The working principle of genotype completion techniques is to infer unobserved variants using common patterns in a reference panel $\mathbf{R}$. This implies that inferred variants are likely to be equal to the most common variants in $\mathbf{R}$. Thus, setting the avatar to such common variants increases the probability that the inferred variants are equal to the avatar ($\Pr(\hat{g}_i = \tilde{g}_i)$), reducing the error in R^y.

Let us denote as $\dot{g}_i$ the most common value in the reference panel for variant i. Depending on the variant's aaf, we have that $\dot{g}_i = 0$, if $\mathsf{aaf} \leq 0.5$, and $\dot{g}_i = 2$ otherwise (variant's values encoded as 1 are never the most common, since they are split in two depending on which of the two chromosomes holds which allele). We compute the *genome avatar for phenotype inference*, denoted as $\tilde{\mathbf{g}}^{\mathbf{g}}$, using a privacy configuration parameter $p_g \in [0,1]$:

$$\tilde{g}_i^g = \begin{cases} \dot{g}_i, & \text{if } g_i = \dot{g}_i, \\ \dot{g}_i, & \text{if } g_i \neq \dot{g}_i, \text{ with probability } p_g, \\ g_i, & \text{if } g_i \neq \dot{g}_i, \text{ with probability } 1 - p_g. \end{cases} \tag{9}$$

Given this creation mechanism, we compute the probability that the adversary succeeds at recovering true genotypes from the avatar considering the two possible avatar values. When $\tilde{g}_i^g \neq \dot{g}_i$, the adversary is certain that the

value observed is the real genotype g_i (third case in (9)), thus succeeds with probability one. On the other hand, when $\tilde{g}_i^g = \dot{g}_i$, the choice that maximizes the adversary's success is to guess that $g_i = \dot{g}_i$. Her success probability is $1 - p_g(1 - \Pr[\dot{g}_i])$, where the second term captures the probability of failure, i.e., $\tilde{g}_i^g = \dot{g}_i$ was a consequence of the second case in (9). Therefore, the privacy level computed as in (8) is:

$$\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{g}} = 1 - \frac{\sum_i \mathbb{1}_{\tilde{g}_i^g \neq \dot{g}_i} + \mathbb{1}_{\tilde{g}_i^g = \dot{g}_i}(1 - p_g(1 - \Pr[\dot{g}_i]))}{n}, \tag{10}$$

Effectively, the parameter p_g balances the privacy and decision precision provided by the avatar. The larger p_g is, the larger the difference between avatar and real genome is (more privacy), but the more different are the decisions with respect to the real genome.

We note that, when related individuals are in the same system, their avatars must be consistent with the Mendelian inheritance laws to avoid inconsistencies when the junction tree algorithm is used for genotype inference of relatives. Given the two parents' avatars generated using the method in (9), we construct offspring avatars by "virtually mating" the parents' avatars. To ensure conservativeness, we choose the most conservative combination that is consistent with the parents for the offspring, instead of choosing at random as happens in reality. Given this creation mechanism, the offspring's avatar is independent from the real offspring genome, and thus does not leak information. It is only related to the parents' avatars which provide the privacy stated in (10).

Membership inference. The membership inference risk is measured in GenoShare as the power of a test establishing whether a statistical summary (e.g., allele frequencies) of a target individual's genome is more similar to the dataset of interest or to the reference population (see Sect. 5.2). Therefore, in order to trigger conservative decisions, an avatar should be more similar to the dataset than the real individual's genotypes.

To build the avatar, we first check which allele contributes more to the dataset aggregate for each variant. Then, with probability p_m, we replace the target's real value with such allele. Formally, we compute the *genome avatar for membership inference*, denoted as $\tilde{\mathbf{g}}^{\mathbf{m}}$ as follows. For each variant i that is in the dataset, given a privacy parameter $p_m \in [0, 1]$:

$$\tilde{g}_i^m = \begin{cases} \max(0, g_i - 1), & \text{with probability } p_m, \text{ if } \mathsf{aaf}_i \geq f_i \\ \min(g_i + 1, 2), & \text{with probability } p_m, \text{ if } f_i > \mathsf{aaf}_i \\ g_i, & \text{with probability } 1 - p_m. \end{cases} \tag{11}$$

The parameter p_m can be used to trade-off privacy and decision precision. Following the same reasoning as for the genome avatar, the success probability of the adversary is 1 when $\mathsf{aaf}_i \geq f_i$ and $\tilde{g}_i^m = 2$, and when $\mathsf{aaf}_i < f_i$ and $\tilde{g}_i^m = 0$. In the former case, she knows that $g_i = 2$, in the latter $g_i = 0$ (third case in (11)). Otherwise, the success of the adversary is $\max(p_m, 1 - p_m)$, depending on the probability of replacement. Then, the term $\Pr[g_i|\tilde{g}_i^m]$ in (8) is:

$$\Pr[g_i|\tilde{g}_i^m] = \begin{cases} 1, & \text{if } \mathsf{aaf}_i \geq f_i \wedge \tilde{g}_i^m = 2 \\ & \vee \mathsf{aaf}_i < f_i \wedge \tilde{g}_i^m = 0 \\ \max(p_m, 1 - p_m), & \text{otherwise .} \end{cases} \tag{12}$$

Kinship inference. Similar to membership inference, the risk of kinship inference depends on the power of a test measuring how similar the genomes of two individuals are (see Sect. 5.3). Essentially, the degree of relationship inferred by this test depends on the amount of overlap weighted by the allele frequency of the sampled variants. Hence, to ensure conservativeness, avatars should be more similar to the target's relative genome that the target itself.

For an individual A, we compute the *genome avatar for kinship inference* with respect to individual B, denoted as $\tilde{\mathbf{g}}^{\mathbf{k}}$, given a privacy parameter $p_k \in [0, 1]$ as:

$$\tilde{g}_i^k = \begin{cases} g_{i,B}, & \text{if } g_i = g_{i,B}, \\ g_{i,B}, & \text{if } g_i \neq g_{i,B}, \text{ with probability } p_k, \\ g_i, & \text{if } g_i \neq g_{i,B}, \text{ with probability } 1 - p_k, \end{cases} \tag{13}$$

where $g_{i,B}$ is B's genotype at variant i.

Since this avatar generation process is analogous to the one for genotype inference, privacy is computed in the same way:

$$\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{k}} = 1 - \frac{\sum_i \mathbb{1}_{\tilde{g}_i \neq g_{i,B}} + \mathbb{1}_{\tilde{g}_i = g_{i,B}}(1 - p_k(1 - \Pr[g_{i,B}]))}{n}, \tag{14}$$

where $\Pr[g_{i,B}]$ is genotype $g_{i,B}$'s prior probability in the population.

Each individual in the database needs to have one genome avatar for kinship inference per relative in the database. When simulating the kinship inference attack, the Attack Engine uses the avatar that corresponds to the closest relative known to the adversary (e.g., because of previous releases).

7.2 Aggregates avatar

Aggregates avatars $\tilde{\mathbf{f}}$ are used when GenoShare receives an aggregates' request $q_m(\mathbf{f}_s)$. As only the membership inference technique makes use of the dataset, one aggregated avatar is sufficient.

Intuitively, conservative decisions for membership-inference attacks should be triggered when the aggregates avatar is more similar to the genomic information to be tested by the adversary than to the population, so individuals would be found to be in the database and GenoShare would prevent the sharing. We construct the *aggregates avatar for membership inference*, denoted as $\tilde{\mathbf{f}}$, as follows. Given privacy parameters $\gamma, p_f \in [0, 1]$, for all i for which $f_i \neq g_i/2$, we sample $\delta_i^{\tilde{f}_i}$ from $\mathcal{U}[0, \gamma|f_i - g_i/2|]$, and generate $\tilde{f}_i$ as follows:

$$\tilde{f}_i = \begin{cases} f_i - \delta_i^{\tilde{f}_i} & \text{if } f_i > g_i/2 \text{ with probability } p_f \\ f_i + \delta_i^{\tilde{f}_i} & \text{if } f_i < g_i/2 \text{ with probability } p_f \\ f_i & \text{with probability} (1 - p_f) . \end{cases} \tag{15}$$

If $f_i = g_i/2$, there is no way to construct a conservative avatar, and the avatar value will take the original value f_i. Therefore, the adversary's success probability is 1 in this (very rare) case. In other cases, the adversary will infer the

frequency that maximizes his success between the original f_i (third case in (15)) and the modified ones (first or second cases in (15)) depending on the parameters. The resulting privacy of this avatar is given as:

$$\mathsf{Priv}_{\tilde{\mathbf{f}}} = \tag{16}$$
$$1 - \frac{1}{n}\sum_i \mathbb{1}_{f_i=g_i/2} + \mathbb{1}_{f_i\neq g_i/2} \max\left(1 - p_f, p_f \frac{\epsilon}{\gamma|f_i - g_i/2|}\right),$$

where the second term in the maximum function is derived from the following. As we rely on the uniform (continuous) distribution to generate $\tilde{f}_i$, the probability of being exactly at f_i is, in general, not defined. Instead, we compute the probability of being in a small interval (represented by ϵ) around f_i, knowing $\tilde{f}_i$:

$$\Pr[|F_i - f_i| \leq \epsilon | \tilde{f}_i] = \begin{cases} 1, & \text{if } \tilde{f}_i = f_i, \\ \frac{\epsilon}{\Delta}, & \text{if } (f_i > g_i/2 \wedge f_i - \Delta \leq \tilde{f}_i < f_i) \\ & \vee (f_i < g_i/2 \wedge f_i < \tilde{f}_i \leq f_i + \Delta), \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

where $\Delta = \gamma|f_i - g_i/2|$. In this case, p_f can be used to trade off privacy and accuracy in a coarse manner, and the parameter γ serves to fine-tune this trade-off. Note that, for the aggregate avatar, contrary to the genome avatar, the privacy value never depends on the specific avatar value $\tilde{f}_i$, mainly because we deal here with continuous values and not three possible discrete values that significantly constrain the space of possible avatars (in conjunction with the requirement to output a conservative query answer).

7.3 Using avatars

The avatar generation mechanisms described above depend on configuration parameters p_x, $x \in \{g, m, k, d\}$, that define the level of privacy provided by the avatars $\mathsf{Priv}_{\tilde{\mathbf{g}}}$, resp. $\mathsf{Priv}_{\tilde{\mathbf{f}}}$. As obtaining a closed expression that expresses the relation between the level of privacy and p_x is extremely complex, deriving analytically configuration values is not possible. However, computing avatars is extremely cheap. Thus, one can efficiently search for adequate parameters (e.g., using the bisection method).

Every time GenoShare is launched, it uses as many avatars as attacks its needs to consider. However, we note that an adversary cannot learn more than a single avatar for a given position by making multiple requests. Indeed, either there is no denial and the adversary learns nothing about any avatar, or there is a denial and this denial will always be based on the same avatar (the most conservative one) for later requests. In other words, there cannot be a denial based on more than one avatar.

8. USING GENOSHARE WITH REAL DATA

We now show how GenoShare can, in practice, support privacy-conscious decisions when sharing genomic data. We consider three use cases in which the adversary makes different requests and has different background knowledge. These use cases are chosen to illustrate how GenoShare reacts to the most likely combinations of requests and background knowledge, and how the interrelations between the attacks influence GenoShare's decision.

8.1 Experimental Setup

Real Data. We run our experiments on the genomes of 351 individuals with ad mixed American ancestries (AMR) from the 1,000 Genomes Project [52]. For each individual, we sample 270k variants across all autosomal chromosomes, and take this to be a representative sample of her genome. We use 250 individuals to build the public reference panel (**R**), and the remaining 101 to simulate the institution's database (**D**). We also build a "sensitive" dataset formed by 50 random individuals in **D** to simulate an HIV-related cohort **H** (any other sensitive disease could be alternatively used here as an example).

GenoShare's Initialization. We set up GenoShare's Attack Engine with the three inference attacks introduced in Section 5. To instantiate the *phenotype inference attack*, we implemented genotype completion using Brian L. Browning's BEAGLE implementation v4 [7, 6], and the junction tree algorithm using the Netica Bayesian network Software [11]. We take the disease-variant associations for the AMR population from the GWAS Catalog [20] for genotype-phenotype mapping. To instantiate *membership* and *kinship* inference attacks, we used our own implementations of the Homer [22] and kinship coefficient [36] inference techniques.

GenoShare's Avatar Engine is set up using the generation techniques in Section 7. For all individuals in **D**, we generate (i) a genome avatar for phenotype inference, (ii) a genome and an aggregates avatar for membership inference, if they are also part of **H**; and (iii) a genome avatar for kinship inference, if they have relatives. Finally, the Risk Measurement Engine is initialized with the risk metrics in Section 6, and we instantiate the Decision Engine with risk thresholds particular to each use case.

8.2 Use Cases

In the following, we assume that the adversary always requests data about a single target individual, or summary statistics about a single disease-related cohort. To consider multiple individuals/cohorts, it suffices to replicate the experiments for all targets.

UC1: Genotype request – no background knowledge on D.
In this scenario, the institution receives consecutive requests for releasing batches of 100 variants of individual A in **D**. We consider that A is part of the HIV-related cohort, **H**, and one of her relatives, B, is also in **D** but not in **H**. A's risk thresholds for phenotype inference risk, membership to the HIV-related cohort inference risk, and kinship inference risk are $\rho_y = 0.9, \rho_m = 0.7, \rho_k = 0.9$, respectively. We consider that the adversary's background knowledge ($\mathcal{B}$) consists of the publicly available reference panel **R**.

Upon reception of a genotype request $q_g(\mathbf{g}_s)$, the institution configures GenoShare with $\mathcal{B}$ and $\boldsymbol{\rho}$ above, and launches it. Then, all the attacks in the Attack Engine are run on A's genome avatars, considering all their interdependencies.

Let us first consider the phenotype inference attack. To illustrate the evolution of A's phenotype inference risk, we consider the adversary's goal is to learn her predisposition to Alzheimer's disease and bipolar disorder. We stress, however, that GenoShare could be configured to consider disclosure of any other genomic-related clinical trait or phenotype. Figure 3 shows this risk's evolution as consecutive batches of variants are released. The solid line represents the risk

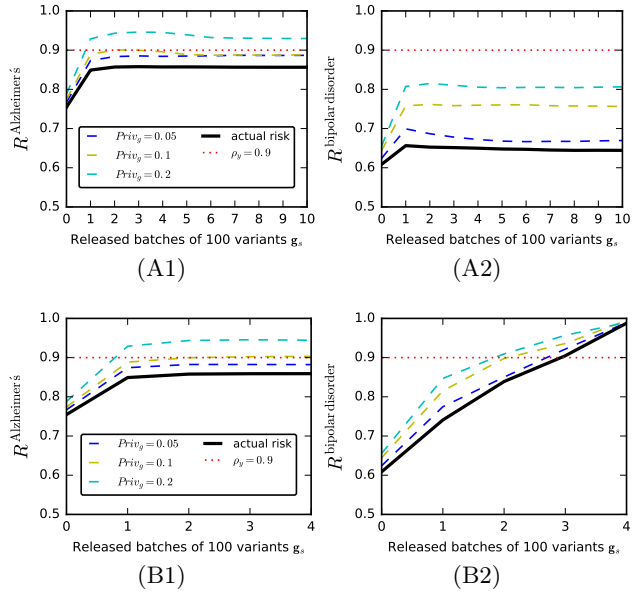


Figure 3: *UC1* – Disease predisposition inference risk, R^y, when releasing A's genotypes in batches of 100: effect of releasing arbitrary variants on the risk of inferring predisposition to Alzheimer's disease (A1) and bipolar disorder (A2); effect of releasing schizophrenia-related variants on the risk of inferring predisposition to Alzheimer's disease (B1), and bipolar disorder (B2).

for A's real genotype, and dashed lines for A's avatars offering privacy $\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{g}} = \{0.05, 0.1, 0.2\}$ (maximum privacy is $\mathsf{Priv}^{max}_{\tilde{\mathbf{g}}\mathbf{g}} = 0.29$). The dotted red line represents A's threshold $\rho_y = 0.9$. The first point in each figure represents the risk before any variant is released, i.e., the prior risk for A computed as in (4) where the adversary's estimation of the target variants' is made according to the alternate allele frequency in the population $\mathsf{aaf}s$.

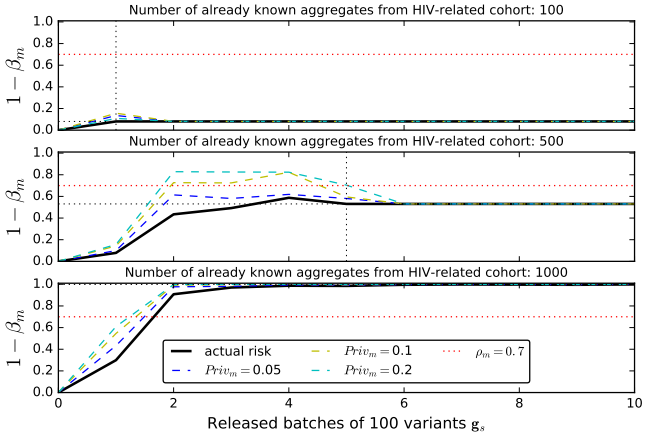


Figure 4: *UC1* – Power of membership inference for different levels of adversarial knowledge ($\alpha_m = 10^{-4}$).

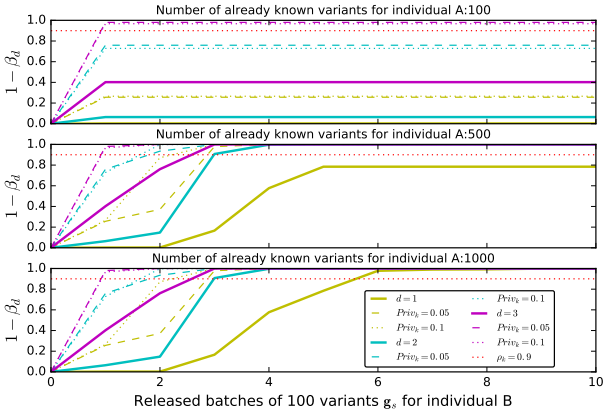
We consider two data-request patterns. The first pattern consists of a series of requests for arbitrary variants, that are not necessarily correlated with any sensitive disease. This case represents the behavior of a researcher looking for new genomic associations with a disease of interest at a genome-wide scale. The results of this experiment are shown in Figure 3(A1-2). First, we observe that released data become part of the adversary's knowledge as $\mathbf{A_g}$, and thus the risk never decreases. Second, as expected, releasing arbitrary (thus likely disease-unrelated) variants slightly affects A's predisposition inference risk for both considered diseases, when computed on the real genotypes. Yet, when avatars are used, as they contain more common variants, this growth is larger. This is because, due to the genotype completion technique that favors the estimation of common values in $\mathbf{R}$, the estimation of the avatar is better.

The second pattern consists of a series of requests for variants correlated with a specific disease, concretely schizophrenia. This represents a typical scenario in which a researcher studies variants of known significance. Results in Figure 3(B1-2) show that releasing these variants does not have a particular impact on the risk of inferring A's predisposition to Alzheimer's disease as these two diseases are genetically uncorrelated [8]. Yet, because of the high genetic correlation between bipolar disorder and schizophrenia, the risk of inferring predisposition to bipolar disorder significantly increases as more schizophrenia-related variants are released.

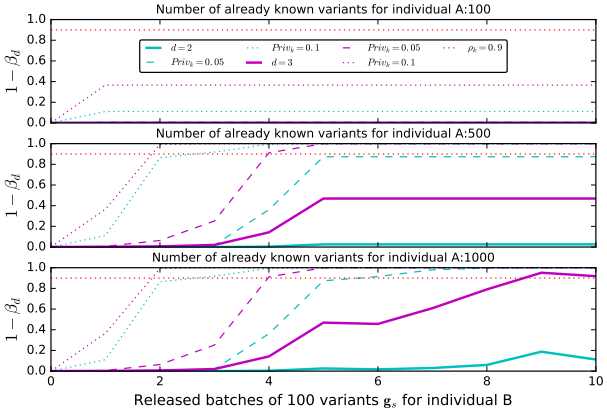
Regarding GenoShare's decision, the larger is $\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{g}}$ the more conservative are the decisions based on the corresponding genome avatars, i.e., for a given inference risk level they allow to disclose less data than those based on the true genome. For instance, let us consider the risk of inferring predisposition to bipolar disorder (Fig. 3(B2)). Given A's threshold, ρ_y, if computations were done on the real genome, GenoShare would allow to release up to 300 genotypes (intersection of the black line and the red dotted line), risking an information leak when the decision is to deny a query. This risk can be mitigated by using avatars, at the cost of releasing less data. The most protective avatar ($\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{g}} = 0.2$, cyan) enables the release of around 200 variants, while the less protective one ($\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{g}} = 0.05$, blue) permits the release of almost as many variants as with the real genome (black).

Given space constraints, in the following, we only show results for the consecutive releases of arbitrary variants. We obtained similar results for variants related to schizophrenia.

Releasing A's variants not only affects her phenotype inference risk but also her membership and kinship inference risks. We show the effect on the risk of inferring her membership to the HIV-related cohort $\mathbf{H}$ for a false positive rate $\alpha_m = 10^{-4}$ in Figure 4. In the different rows of Figure 4, we consider that the adversary has obtained an increasing number of aggregates from previous queries ($\mathbf{A_m}$) to be incorporated to his background knowledge $\mathcal{B}$. Unsurprisingly, the more genotypes are revealed, the stronger is the inference power, up to the point where the number of genotypes and aggregate statistics released are the same (vertical dotted line). Then the inference power remains constant because there are no more aggregate statistics to gain information from (horizontal dotted line). We observe that genotype completion helps membership inference by increasing the inference power before the vertical dotted line. Indeed, thanks to genotype intra-correlations, a larger number of genotypes than those made available to the adversary can be used in the attack (see Fig. 10 in Appendix C(A) for comparison with the membership inference power without the contribution of genotype completion). The behavior of the avatars (dashed lines) is similar to the previous case, privacy can



(A) First-degree relationship between A and B



(B) Second degree relationship between A and B

Figure 5: *UC1* – Power of kinship inference for different degrees ($\alpha_d = 10^{-4}$).

be enhanced at the cost of releasing less information. We note that, the more variants are released, the less different is the avatar from the actual genotype. Hence, the inference power based on the avatar and real genome converges.

Finally, Fig. 5 shows the effect of genotype release on risk of inferring A and B's kinship for different degrees of relatedness $d = \{1, 2, 3\}$ and a false-positive rate $\alpha_d = 10^{-4}$. Similarly to the previous case, each row assumes that the adversary has an increasing number of variants from B as part of his knowledge $\mathbf{A_g}$. We consider two cases where A and B are first and second degree relatives. In the first case (Fig. 5(A)), we see that the kinship inference power is already maximized when 300 variants are disclosed for both individuals. Because of the closed test procedure, the power for $d > 1$ is also maximized. However, the kinship inference power for $d = 0$ (monozygotic twins) is negligible regardless of the number of variants used in the attack, as there is no possibility that this is the case. In the second case (Fig. 5(B)), few variants suffice to infer with significant power that the individuals are at least third degree relatives ($d = 3$). Yet, 1,000 variants are not enough to determine their real kinship, $d = 2$, with high certainty. In fact, up to 4,000 variants are necessary to maximize this inference power (see Fig. 11 in Appendix C). Again avatars perform as expected, enabling a trade-off between amount of released data and privacy in case of query denials.

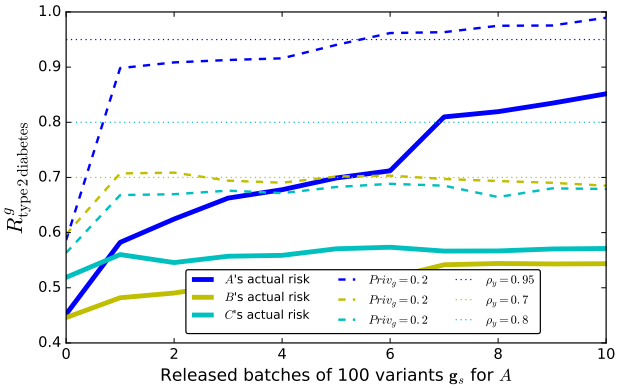
We recall that GenoShare denies a query as soon as any of the inference risks breaches its corresponding threshold. For instance, if we consider an avatar with $\mathsf{Priv}_{\tilde{\mathbf{g}}\mathbf{s}} = 0.05$, and adversarial background knowledge of 500 aggregates from **H** and 500 genotypes from B (second degree relative), GenoShare prevents the release of data at the fourth request because of the kinship inference risk being too high.

UC2: Genotype request – kinship background knowledge.
In this scenario, the institution holding **D** receives consecutive requests for releasing genotypes of arbitrary pathogenic variants (i.e., related to a disease) of individual A. A, and also her parents B and C, are part of the HIV-cohort **H**. Their degree of kinship is already known by the adversary and it is part of his background knowledge $\mathcal{B}$. For example, he could have inferred this information through the kinship inference attack when A's, B's and C's genotypes were released by GenoShare , or if their kinship information was publicly available on Facebook [25]. Moreover, we assume that A, B, and C have different privacy concerns: A is not worried about any potential privacy breach and sets very permissive thresholds $\boldsymbol{\rho}_A = \{\rho_y = \rho_m = 0.95\}$, whereas B and C have more restrictive preferences $\boldsymbol{\rho}_B = \{\rho_y = \rho_m = 0.7\}$ and $\boldsymbol{\rho}_C = \{\rho_y = \rho_m = 0.8\}$ (Note that, as kinship is known, we disregard the kinship threshold.)

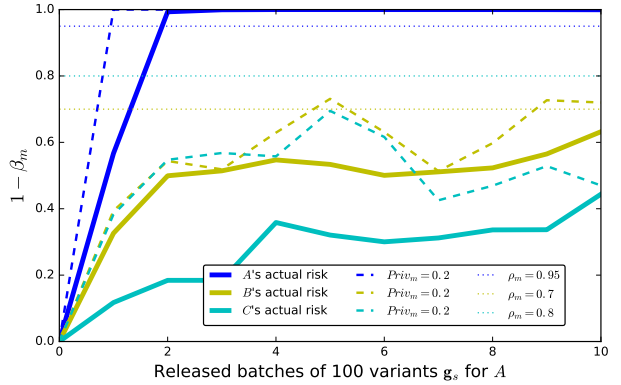
Upon reception of a genotype request $q_g(\mathbf{g}_s)$, the institution configures GenoShare with $\mathcal{B}$, $\boldsymbol{\rho}_A$, $\boldsymbol{\rho}_B$ and $\boldsymbol{\rho}_C$, and launches it. GenoShare runs both phenotype and membership inference considering their interrelation, and computes the risk of a privacy breach for the three individuals. Figure 6(A) illustrates the evolution of A's, B's, and C's predisposition to type 2 diabetes inference risk. Because of the kinship effect, releasing A's genotypes has an effect on B's and C's risk computed with both the real genotype (solid) and avatars (dashed). We note that, because the three individuals are involved, it is enough that at least one of their risk thresholds is exceeded to deny the request. For instance, if avatars are used, even though the first query would be deemed safe for A (blue), it would be denied because it implies that the risk for B (yellow) goes above her threshold.

Because it improves genotype completion, kinship information also significantly affects membership inference. Figure 6(B) illustrates the evolution of the risk of membership to the HIV-related cohort **H** inference for A, B, and C for $\alpha_m = 10^{-4}$ when 1,000 aggregates are already known to the adversary. Similarly to the previous use case, we observe how releasing A's genotypes also increases the membership inference risk for B and C. Also, observe that, as in the third row of Figure 4, genotype completion helps the adversary by significantly increasing his inference power (even if only 100 genotypes are available after the first query, the adversary can exploit the 1000 aggregates that are available).

UC3: Aggregate request – no background knowledge on D.
In this last scenario, the institution holding **D** receives consecutive requests for releasing aggregate statistics $\mathbf{f}_s$ of the HIV-related cohort **H**. We consider the same privacy preferences and background knowledge as in UC1. Upon reception of a request for aggregate data, the institution configures GenoShare with public background knowledge and launches it. In this particular case, GenoShare only runs the membership inference attack because it is the only attack for which obtaining new aggregates helps the adversary.



(A) Risk of phenotype inference for type 2 diabetes



(B) Risk of membership inference ($\alpha_m = 10^{-4}$) when the adversary already knows 1,000 aggregates from the HIV-related cohort

Figure 6: *UC2* – Disease predisposition and membership inference risk for A, B, and C

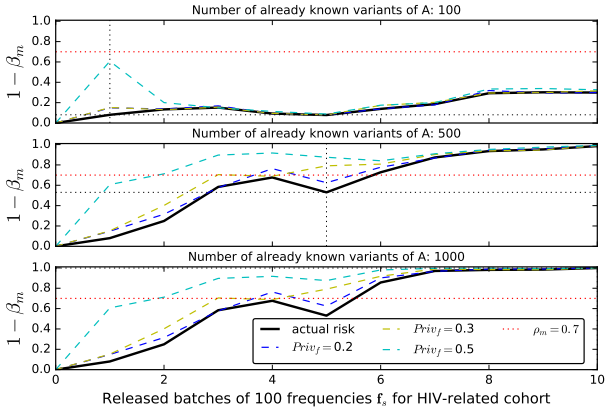


Figure 7: *UC3* – Effect of releasing batches of 100 aggregates on the risk of membership inference ($\alpha_m = 10^{-4}$, $\gamma = 0.25$).

Figure 7 shows the effect of a series of aggregate data releases on the risk of inferring A's membership to the HIV-related cohort $\mathbf{H}$, for a false-positive rate $\alpha_m = 10^{-4}$ and different amount of A's variants available to the adversary. Unsurprisingly, the more aggregates are released, the higher the inference power. It is important to note that, thanks to the *genotype completion* carried out in the phenotype inference, the adversary can use every new released aggregated frequencies even if the corresponding genotypes of A have not yet been revealed (i.e., after the black dotted lines cross). Thus, the risk of membership inference keeps growing with every extra frequency observed by the adversary. Without the contribution of the genotype completion, inference power stays constant since additional aggregate statistics cannot be used by the membership attack (see Figure 10(B) in Appendix C). As in previous cases, we observe how avatars enable to trade off privacy for amount of data released.

Finally, this use case also shows GenoShare's utility in the case where the adversarial background knowledge consists of the full genome of some individuals in $\mathbf{H}$. For example, individuals that put their genome online (e.g., on OpenSNP), or have been sequenced at a direct-to-customer genomic service (e.g., 23andMe) that keeps a copy of their genome. In these circumstances, as the genome is already known, the only attack that GenoShare can mitigate is the membership inference attack based on aggregate data requests. We emphasize that, for GenoShare to not underestimate the membership inference risk in such a case, it is the responsibility of the individuals to communicate to the institution holding their genomic data that other copies are available elsewhere. In this case, GenoShare's background knowledge should be set to account for the adversary's knowledge of the complete target's genome before its execution.

8.3 GenoShare's Performance

GenoShare is not intended to be a real time tool, but to be run offline. We conduct performance measurements on an 8-core Intel Xeon CPU E3-1270 V2 processor 3.50GHz, and 32 GB of memory, running Debian Linux. We measure the time required for computing the inference risks for the three considered attacks by executing GenoShare 10 times on 10 different requests of 100 arbitrary variants, and report the average over the 100 experiments. Although the considered inference techniques are naturally parallelizable, our measurements are made on a single thread of execution and thus represent an upper bound for computation time.

We find that, despite its apparent complexity, the use of GenoShare entails very reasonable overhead. As expected, since they only require fast arithmetic operations, the computation time of both the kinship and membership inference attacks is within the second hence negligible, and grows linearly with the number of variants to be tested. Similarly, the generation of avatars, which is performed only once during GenoShare's initialization, is also negligible. On the other hand, the phenotype inference attack, that uses BEAGLE's Java implementation for *genotype completion*, takes on average around 8 minutes and 14 seconds at each new request. We note that this timing is strongly influenced by our choice of running BEAGLE with default parameters and limiting the amount of memory allocated for the Java virtual machine to 2GB, and could be reduced by optimizing the configuration (e.g., with a typical 22-node cluster it requires less than 23 seconds). We also note that the genotype completion computation time depends both on the number of variants to be considered across the genome (in our case 270k variants), and on the number of individuals in the reference panel $\mathbf{R}$ (in our case 250 individuals). We refer the reader to the original publication for more details on how BEAGLE scales [7, 6].

In terms of storage, for each individual in the database, GenoShare requires one avatar per kind of attack instanti-

ated. The size of an avatar is at most the size of the set of human genetic variations, i.e., roughly 1% of a full genome. Thus, GenoShare's storage requirements are certainly practical.

9. RELATED WORK

We revise attacks and defense mechanisms for genomic privacy which are most relevant to our work. We refer the reader to existing surveys [13, 40] for an exhaustive review of the state of the art.

Attacks. A first group of relevant attacks, referred to as *attribute disclosure*, can be subdivided into two categories. The first category, attribute disclosure via genomic completion, includes all techniques that rely on the intra- and inter-genome correlations to infer unobserved variants [34] to reduce genomic privacy of individuals [46] and relatives [24]. In this work, we improve previous inference techniques by taking both familial correlations and intra-genome correlations using *genotype completion*. The second category, attribute disclosure via membership inference, exploits knowledge of summary statistics of a given dataset to infer that a known genome is part of it [10, 22, 27, 47, 57, 59, 64]. As such datasets are typically associated to a disease of interest, inferring membership unveils very sensitive attributes. Commonly exploited statistics are allele frequency and genotype counts, or statistics about linkage disequilibrium. In this work, we explore for the first time the effectiveness of these attacks in presence of incomplete information, and the benefits of genotype completion on membership inference.

A second group of relevant papers deals with *kinship inference*. Previous works show how to exploit inter-genome correlations to infer familial relationships based on the amount of genomic data shared by individuals in genome-wide association studies with distinct subpopulations [36] or in admixed populations [54]. Recently, Arthur et al. have proposed a toolkit for analyzing large cohorts of whole-genome sequenced samples that includes kinship inference relying on state-of-the-art methods [1]. Our contribution is the framing of kinship inference as from a privacy perspective, both as an attack to reveal familial relationship and as complement to genotype completion to increase the amount of genetic information that can be inferred about an individual.

Finally, we revise *re-identification attacks* that de-anonymize genomic data by relying on auxiliary knowledge. Gymrek et al. have proposed an attack of this kind [18], which uses short tandem repeats on the Y chromosome (Y-STRs) to map anonymous genomes to those available in a recreational genetic genealogy database online to recover the surname of their (male) owner. Similarly, Humbert et al. link de-identified genotypes to online profiles, such as those from online social networks, by relying on genomic variants that influence phenotypic traits [26]. In this paper, we do not consider these attacks for space constraints and because they highly rely on the access to external background knowledge that may be difficult to access (e.g., the Y-STR database used by Gymrek et al. has been put offline after this attack was made public). Yet, we note that GenoShare's Attack Engine can easily accommodate re-identification attacks by integrating it in the different engines.

Protection mechanisms. In addition to de-identification, which is a necessary but not sufficient protection method as shown above, there exist two approaches for genomic privacy protection. The main idea behind the first approach is to properly apply noise, e.g., achieving differential privacy, on summary statistics for protecting a study participants' privacy and thus thwarting attribute disclosure attacks [28, 49, 55, 63]. Fredrikson et al. show that, however, using differential privacy in pharmacogenetics can lead to an unacceptable loss of utility, e.g., exposure of patients to an increased risk of stroke, bleeding events, and mortality [16]. Also, practitioners require exact genomic data to avoid false genotype-phenotype associations. Bhaskar et al. propose a noiseless version of differential privacy but their solution makes some statistical assumptions on the data that are too restrictive for the genomics setting [4]. Our solution does not perturb the released aggregated data at all, regardless of the data distribution. Moreover, the aforementioned protection mechanisms are not suitable for the release of an individual's variants. The second approach relies on secure storage and processing [3, 12, 23, 40, 60, 39] which are complementary to our proposed solution. In fact, secure processing protects only the data while processing, but GenoShare considers all the disclosed/shared information, including also the results of a computation.

10. CONCLUSION

Academic solutions for privacy-preserving sharing of genomic data have mostly focused on data perturbation. Such solutions, however, damage the utility of the data and thus have not been accepted by practitioners. In this work, we introduce GenoShare, a framework that supports privacy-informed decision-making when sharing genomic data. GenoShare quantifies the risk of sensitive attribute disclosure, and prevents the automatic sharing of data if the risk is deemed too high with respect to privacy thresholds encoded using novel meaningful sensitive attribute-oriented metrics. Otherwise, it releases *exact data* as requested by genomics research practitioners. Furthermore, GenoShare implements avatar genomes to protect individuals' real genotypes from inferences stemming from query denials.

To the best of our knowledge, GenoShare is the first framework to jointly consider relevant attacks in genomic privacy in presence of incomplete information. It provides a principled answer to the privacy concerns that have plagued the genomic community for the last decade, and thus it is a firm step forward to enable the responsible and privacy-respecting use of genomic data in research and medical environments. We hope that it will dramatically improve the current situation in institutions, thereby accelerating the slow and costly processes carried out by committees and lawyers by serving as support for more informed decisions.

Although we have focused on the protection of genomic data, the systematic principles underlying GenoShare make it suitable to deal with other data types where correlation with sensitive information can be detrimental for an individual's privacy, e.g., other 'omics' data such as transcriptomic. Furthermore, GenoShare can also be used to understand the risk incurred when voluntarily disclosing information to find others who have a similar rare disease and share experiences as on PatientsLikeMe [43] or to safely enjoy recreational genomics or direct-to-consumer genomics services.

11. REFERENCES

[1] R. Arthur, O. Schulz-Trieglaff, A. J. Cox, and J. M. O'Connell. Akt: Ancestry and kinship toolkit. *Bioinformatics*, 2016.

[2] M. Backes, P. Berrang, A. Hecksteden, M. Humbert, A. Keller, and T. Meyer. Privacy in epigenetics: Temporal linkability of microrna expression profiles. In T. Holz and S. Savage, editors, *25th USENIX Security Symposium*, pages 1223–1240. USENIX Association, 2016.

[3] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. In *ACM Conference on Computer and Communications Security, (CCS)*, pages 691–702, 2011.

[4] R. Bhaskar, A. Bhowmick, V. Goyal, S. Laxman, and A. Thakurta. Noiseless database privacy. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 215–232. Springer, 2011.

[5] V. Bindschaedler, R. Shokri, and C. A. Gunter. Plausible deniability for privacy-preserving data synthesis. *PVLDB*, 10(5):481–492, 2017.

[6] B. L. Browning and S. R. Browning. Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126, 2016.

[7] S. R. Browning and B. L. Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[8] B. Bulik-Sullivan, H. K. Finucane, V. Anttila, A. Gusev, F. R. Day, P.-R. Loh, L. Duncan, J. R. Perry, N. Patterson, E. B. Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 2015.

[9] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, et al. Modeling 3d facial shape from dna. *PLoS Genet*, 10(3):e1004224, 2014.

[10] D. Clayton. On inferring presence of an individual in a mixture: a bayesian approach. *Biostatistics*, page kxq035, 2010.

[11] N. S. Corp. Netica. `https://www.norsys.com/index.html`, 2017. Last Accessed: August 30, 2017.

[12] G. Danezis and E. De Cristofaro. Fast and private genomic testing for disease susceptibility. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014.

[13] Y. Erlich and A. Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15(6):409–421, 2014.

[14] Facebook. `https://www.facebook.com/`, 2017. Last Accessed: August 30, 2017.

[15] S. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. In *ICDM*, pages 628–635, 2011.

[16] M. Fredriksen, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium*, 2014.

[17] D. Greenbaum, A. Sboner, X. J. Mu, and M. Gerstein. Genomics and privacy: Implications of the new reality of closed data for the field. *PLoS Comput Biol*, 7(12), 12 2011.

[18] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[19] A. Harmanci and M. Gerstein. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods*, 13(3):251–256, 2016.

[20] L. A. Hindorff, H. A. Junkins, P. Hall, J. Mehta, and T. Manolio. A catalog of published genome-wide association studies. `http://www.genome.gov/gwastudies`, 2011. Last Accessed: August 30, 2017.

[21] HIPAA News. 480,000 Patients Notified of Radiology Regional Center PHI Exposure. `http://www.hipaajournal.com/480000-patients-notified-of-radiology-regional-center-phi-exposure-8322/`.

[22] 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.

[23] Z. Huang, E. Ayday, J. Fellay, J. Hubaux, and A. Juels. Genoguard: Protecting genomic data against brute-force attacks. In *IEEE Symposium on Security and Privacy*, pages 447–462. IEEE Computer Society, 2015.

[24] M. Humbert, E. Ayday, J. Hubaux, and A. Telenti. Reconciling utility with privacy in genomics. In *13th Workshop on Privacy in the Electronic Society (WPES14)*, pages 11–20. ACM, 2014.

[25] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the lacks family: Quantification of kin genomic privacy. In *ACM Conference on Computer and Communications Security, (CCS)*, 2013.

[26] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux. De-anonymizing genomic databases using phenotypic traits. *Proceedings on Privacy Enhancing Technologies*, 2015(2):99–114, 2015.

[27] K. B. Jacobs, M. Yeager, S. Wacholder, D. Craig, P. Kraft, D. J. Hunter, J. Paschal, T. A. Manolio, M. Tucker, R. N. Hoover, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41(11):1253–1257, 2009.

[28] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *KDD*, pages 1079–1087, 2013.

[29] K. Kenthapadi, N. Mishra, and K. Nissim. Simulatable auditing. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 118–127. ACM, 2005.

[30] K. Kenthapadi, N. Mishra, and K. Nissim. Denials leak information: Simulatable auditing. *Journal of*

Computer and System Sciences, 79(8):1322–1340, 2013.

[31] H. Kim, E. Bell, J. Kim, A. Sitapati, J. Ramsdell, C. Farcas, D. Friedman, S. F. Feupe, and L. Ohno-Machado. iconcur: informed consent for clinical data and bio-sample use for research. *Journal of the American Medical Informatics Association*, 2016.

[32] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

[33] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[34] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[35] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[36] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

[37] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7):906–913, 2007.

[38] R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[39] P. J. McLaren, J. L. Raisaro, M. Aouri, M. Rotger, E. Ayday, I. Bartha, M. B. Delgado, Y. Vallet, H. F. Günthard, M. Cavassini, et al. Privacy-preserving genomic testing in the clinic: a model using HIV treatment. *Official journal of the American College of Medical Genetics and Genomics*, 2016.

[40] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):6, 2015.

[41] A. Nowogrodzki. Spiking genomic databases with misinformation could protect patient privacy. *Nature News*, 2016.

[42] OpenSNP. `https://opensnp.org/`, 2017. Last Accessed: August 30, 2017.

[43] PatientsLikeMe. `https://www.patientslikeme.com`, 2017. Last Accessed: August 30, 2017.

[44] Premier Healthcare. Notice to Our Patients Regarding a Security Incident. `http://www.premierhealthcare.org/incident-2016-03.html`.

[45] L. R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., 1990.

[46] S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.-P. Hubaux, and Z. Kutalik. Quantifying genomic privacy via inference attack with high-order snv correlations. In *2nd International Workshop on Genome Privacy and Security (in conjunction with IEEE S&P; 2015)*, 2015.

[47] 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.

[48] S. S. Shringarpure and C. D. Bustamante. Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, 2015.

[49] S. Simmons and B. Berger. Realizing privacy preserving genome-wide association studies. *Bioinformatics*, 32(9):1293–1300, 2016.

[50] G.-W. A. Studies. `http://www.genome.gov/20019523`, 2017. Last Accessed: August 30, 2017.

[51] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[52] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[53] The Wall Street Journal. Anthem: Hacked Database Included 78.8 Million People. `http://www.wsj.com/articles/anthem-hacked-database-included-78-8-million-people-1424807364`.

[54] T. Thornton, H. Tang, T. J. Hoffmann, H. M. Ochs-Balcom, B. Caan, and N. Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012.

[55] C. Uhler, A. Slavkovic, and S. E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

[56] U.S. Department of Health and Human Services . Breach portal: Notice to the secretary of hhs breach of unsecured protected health information. `https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf`. Last Accessed: August 30, 2017.

[57] P. M. Visscher and W. G. Hill. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genetics*, 5(10):e1000628, 2009.

[58] I. Wagner. Evaluating the strength of genomic privacy metrics. *ACM Transactions on Privacy and Security (TOPS)*, 20(1):2, 2017.

[59] 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 *ACM Conference on Computer and Communications Security, (CCS)*, pages 534–544, 2009.

[60] R. Wang, X. Wang, Z. Li, H. Tang, M. K. Reiter, and Z. Dong. Privacy-preserving genomic computation through program specialization. In *ACM Conference on Computer and Communications Security, (CCS)*, pages 338–347, 2009.

[61] X. Xiao and Y. Tao. Dynamic anonymization:

accurate statistical analysis with privacy preservation. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 107–120. ACM, 2008.

[62] F. Yu, S. E. Fienberg, A. B. Slavković, and C. Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 2014.

[63] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *International Conference on Privacy in Statistical Databases*, pages 170–184. Springer, 2014.

[64] 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.

APPENDIX

A. GENOTYPE IMPUTATION

Inference for target's unobserved variants. As explained in Section 5.1, in order to infer genotypes of unobserved variants for the target individual given the genotypes of some observed variants and the genotypes of individuals in a public reference panel, we use a well-established statistical technique called *genotype imputation* [37]. In particular, we compute the posterior probabilities $Pr(\mathbf{g}_u|\mathbf{g}_o, R)$, where $\mathbf{g}_u$ is the set of unobserved variants, $\mathbf{g}_o := \mathbf{A}_\mathbf{g} \cup \mathbf{g}_s \cup \mathcal{B}$ is the set of observed variants within the same genome, and R is the genotype information for the reference individuals whose set of observed variants is $\mathbf{g} := \mathbf{g}_u \cup \mathbf{g}_o$.

At a high level, genotype imputation works by using patterns of blocks of highly correlated variants, so-called *haplotypes*, in the reference panel to predict unobserved variants when only a subset $\mathbf{g}_o$ of $\mathbf{g}$ has been observed. By definition, a haplotype is a set of variants on a chromosome that tend to always occur together, i.e., that are statistically correlated. This process is illustratged in Figure 8.

Let us call "reference haplotypes" the set of haplotypes in the panel of a reference individual, and "target haplotypes" the set of observed haplotypes in the target genome. A target genome can be considered as a mosaic of reference haplotypes. If a reference haplotype is similar to the target genotype in the region of a given target haplotype, then such a reference haplotype can be chosen to be the one in the mosaic at that target observed haplotype position.

Now, more formally, let H_R be the set of reference haplotypes in the panel of reference individuals, and let H_T be the set of the target's observed haplotypes, in chromosome order. Then, the target's genome can be modeled as a Hidden Markov Model (HMM) where the state space is the set of all ordered pairs (h_T, h_R) whose first element $h_T \in H_T$ is a target's observed haplotype and whose second element $h_R \in H_R$ is a reference haplotype [34]. When modeling a target genotype, a state (h_T, h_R) has high probability if the target genotype is well represented by a mosaic of reference haplotypes. Let us denote the set of model states at haplotype $h_T \in H_T$ as $H_{h_T} = \{(h_T, h_R) : h_R \in H_R\}$. Then, we can use the HMM forward-backward algorithm [45] to estimate the probabilities $Pr(S_{h_T} = h)$, $S_{h_T} \in H_{h_T}$ conditional on the HMM model and the values of the observed variants on the target haplotype. Li and Stephens model [34] is the state-of-the-art model for summarizing and interpreting genetic intra-correlations, i.e., *Linkage Disequilibrium (LD)* among multiple variants by considering all loci simultaneously, rather than pairwise. We refer the reader to the original paper for further details.

Inference for relatives' unobserved variants. We make use of a well-established inference algorithm, the *junction tree (or clique tree)* algorithm [32] to infer the hidden variants in the relatives' genomes from the observed variants $\mathbf{g}_o$ and variants imputed in the target individual $\hat{\mathbf{g}}$. As propagation of the evidence has already been carried out throughout the genome with genotype imputation (horizontal inference), The genotype inference for unobserved variants of target's relatives focuses on familial correlations (i.e., across relatives). Thus, without loss of generality, we focus our description on a single variant.

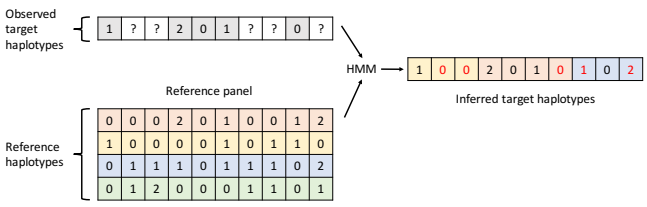


Figure 8: High-level representation of the genotype imputation technique.

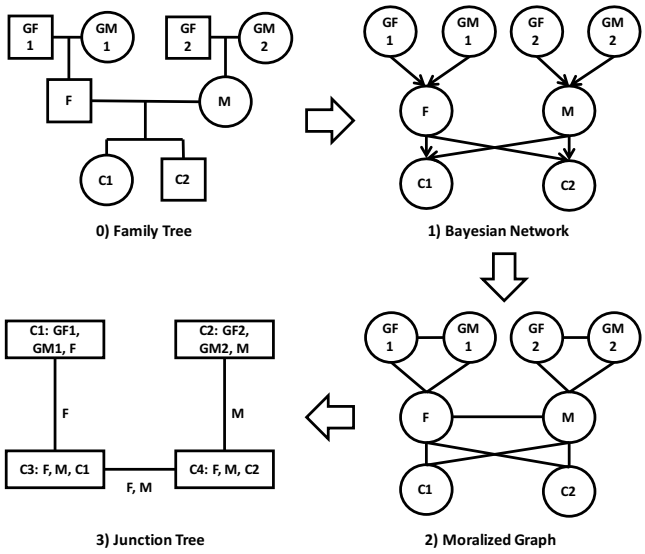


Figure 9: High-level representation of the junction tree algorithm: From a family tree (0), a Bayesian network is constructed (1) and then moralized through triangulation in order to obtain a moralized graph (2). Finally, The moralized graph is transformed into a clique tree (or junction tree) (3) by guaranteeing that for each pair of cliques U, V with intersection S, all cliques on path between U and V contain S.

Because probabilistic inference induces the computation of marginal posterior probabilities, its computational complexity increases exponentially in the number of considered variables, i.e., relatives' variants, if the marginalization is carried out directly on the global joint distribution between these variables. As already mentioned in 5.1, the idea behind the junction tree algorithm and its core inference algorithm, belief propagation, is to split the global joint probability distribution that represents the genomic variants of r relatives at a given variant v into smaller distributions that keep only dependent variants (variables) together.

In our context, thanks to the Mendelian inheritance laws, we can split the global joint distribution into smaller probability functions, as follows:

$$\Pr(g_{v,1}, g_{v,2}, \ldots, g_{v,r}) = \Big(\prod_{j \in \text{founders}} \Pr(g_{v,j})\Big)\Big(\prod_{k \notin \text{founders}} \Pr(g_{v,k}|g_{v,m(k)}, g_{v,f(k)})\Big), \tag{18}$$

in order to use allele probability distributions, where *founders* is the set of individuals who have no parents (with observed genotype data) in the family tree of interest, and $m(k)$ and

$f(k)$ are the mother and father of k. As shown in Figure 9, from this factorization, we can construct a Bayesian network whose nodes represent the variants of the r relatives of interest and whose (directed) edges represent direct dependencies between them. In this Bayesian network, each child node (i.e., not founder) have two parent nodes, as in the real (biological) life, and it is defined by a conditional probability table (representing $\Pr(g_{v,k}|g_{v,m(k)}, g_{v,f(k)})$) that is given by the Mendelian inheritance probabilities.

The only issue that could lead to approximate inference is the fact that siblings in a real family generate (undirected) loops in the underlying Bayesian network. The junction tree algorithm removes these loops by clustering each child node with their two parent nodes when the two parents have more than one offspring, as shown in Fig.9. Then, the belief propagation algorithm computes the marginal probabilities of each clique separately, and propagates its computation to other cliques in the tree. Due to the tree structure of the underlying graph, the inference algorithm converges in only two iterations, one forward and one backward, and is linear in the number of relatives r, and in the number of variants $|V|$ (since the inference at each variant can be run independently from another). Note that the algorithm is, in general, exponential in the maximal clique size, but that this size being equal to 3 in our case, it becomes negligible compared to the number of relatives and variants.

B. KINSHIP INFERENCE CRITERIA

Table 1 shows the inference criteria for different degrees of kinship according Manichaikul et al. [36].

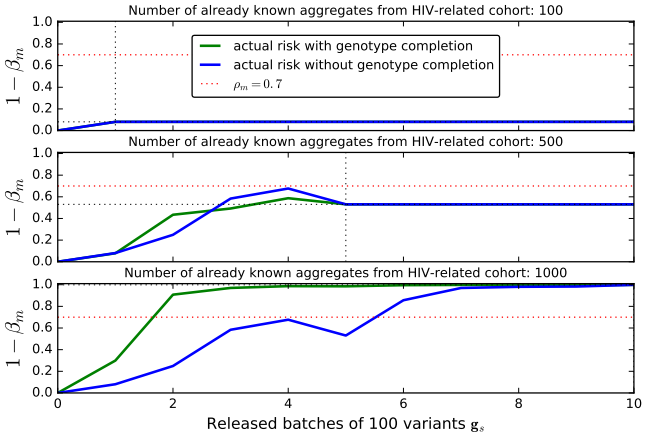
C. SUPPLEMENTARY FIGURES

In this appendix we provide extra figures to enable a better understanding of the effect of the *genotype completion* on the membership inference attack and of genotype data releases beyond 1,000 variants on the kinship inference attack for second-degree relatives.

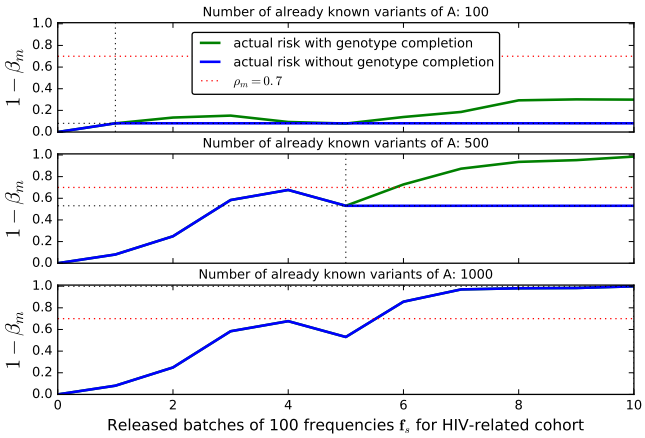
Membership inference with or without genotype completion. We show in Figure 10(A) the effect that genotype completion has on the membership inference power for the scenario described in the first use case of Section 8.2, in which aggregate data of some variants of individuals in the HIV-related cohort **H** have already been revealed and queries request genotypes of the target individual. We observe that, after the vertical dotted line (which represents equal number of aggregated statistics and genotypes available to the adversary), when genotype completion is not used (blue line), there is no difference with respect to the case when genotype completion is used (green line). This is because, regardless of the number of inferred genotypes, the adversary does not have access to more aggregated statistics. However, we observe (especially in the third row) that before this line, the effect of genotype completion of the adversary's inference power is significant as he can use more genotype values than those released by the queries.

Relationship	$\phi_{A,B}$	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{3/2}}$
Parent-offspring	$\frac{1}{4}$	$\left[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}\right]$
Full siblings	$\frac{1}{4}$	$\left[\frac{1}{2^{5/2}}, \frac{1}{2^{3/2}}\right]$
2nd degree	$\frac{1}{8}$	$\left[\frac{1}{2^{7/2}}, \frac{1}{2^{5/2}}\right]$
3rd degree	$\frac{1}{16}$	$\left[\frac{1}{2^{9/2}}, \frac{1}{2^{7/2}}\right]$
Unrelated	0	$< \frac{1}{2^{9/2}}$

Table 1: Kinship inference criteria based on the estimated kinship coefficient $\phi_{A,B}$.



(A) Releasing genotypes g_i when aggregates f_i are known



(B) Releasing aggregates f_i when genotypes g_i are known

Figure 10: Power of membership inference for $\alpha_m = 10^{-4}$: interleaving genomic and aggregated queries

Similarly, Figure 10(B) shows the effect of genotype completion for the scenario described in the third use case of Section 8.2, in which the genotypes of some variants of the target individual are already known and the system releases aggregated statistics for the same variants. We see that in this case, before the vertical dotted line, genotype completion has no impact on the inference power as the adversary can only use the aggregate statistics that are revealed for the attack. Yet, after the vertical dotted line, the adversary can increase his inference power with every extra aggregate statistic he observers thanks to genotype completion.

Kinship inference. Figure 11 complements Figure 5 by showing the evolution of kinship inference power beyond the release of 1,000 genotypes for the target individual A with respect to a second-degree relative B. As mentioned above,

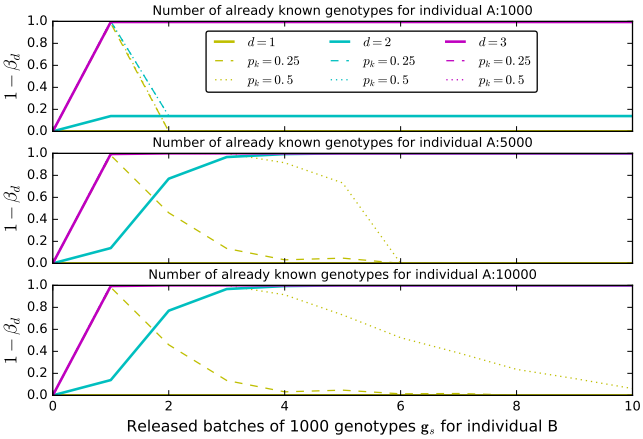


Figure 11: Power of kinship inference for different degrees and $\alpha_d = 10^{-4}$ in the case of second degree relationship between A and B.

only after about the genotypes of 4,000 variants are released, the adversary can maximize her power of kinship inference for $d = 2$.

D. NOTATION TABLE

For reference, Table 2 summarizes the notation we use throughout this paper.

Notation	Description
aaf_i	Alternate allele frequency at position i
y	Disease susceptibility to be protected
$\Psi(y)$	Set of variants associated with disease y
$\omega^y = (\omega_1^y, \ldots, \omega_n^y)$	Set of effect-size coefficients for variants in $\Psi(y)$
$\phi_{A,B}$	Kinship coefficient for individual A and B
$\mathbf{g} = (g_1, \ldots, g_n)$	Set of genotypes for a real genome
$\mathbf{g_o}, \mathbf{g_u}$	Set of observed, unobserved genotypes used by genotype inference
R	Panel of reference individuals used by genotype inference
$\hat{\mathbf{g}} = (\hat{g}_1, \ldots, \hat{g}_n)$	Set of inferred genotypes
$\tilde{\mathbf{g}} = (\tilde{g}_1, \ldots, \tilde{g}_n)$	Set of genotypes for an avatar genome
$\dot{\mathbf{g}} = (\dot{g}_1, \ldots, \dot{g}_n)$	Set of most common genotypes in the population
$\mathbf{f} = (f_1, \ldots, f_n)$	Set of aggregated statistics for a real database
$\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_n)$	Set of aggregated statistics for an aggregated avatar
$q_g(\mathbf{g}_s)$	Query for the genotypes of a subset s of variants
$q_m(\mathbf{f}_s)$	Query for the aggregated statistics on a subset s of variants
$\mathbf{A_g}$	Genotypes revealed in previous queries
$\mathbf{A_m}$	Aggregated statistics revealed in previous queries
R^y	Risk of disclosing predisposition to disease y through phenotype inference attack
R^m	Risk of disclosing database membership through membership inference attack
R_d^k	Risk of disclosing familial relationship of degree d through kinship inference attack
α_m, β_m	Type I and II errors for database membership inference
α_d, β_d	Type I and II errors for kinship inference of degree d
$\boldsymbol{\rho}$	Set of thresholds on inference risks
ρ_y	Threshold on the risk of disclosing predisposition to disease y
ρ_m	Threshold on the risk of disclosing membership m to the database
ρ_k	Threshold on the risk of disclosing kinship
$\mathcal{B}$	Auxiliary information available to the adversary
$\mathsf{Priv}_{\tilde{\mathbf{g}}}$	Genome avatar's privacy
$\mathsf{Priv}_{\tilde{\mathbf{f}}}$	Aggregated avatar's privacy

Table 2: Notation used throughout the paper. For all variables, bold indicates a set.