

Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores

Brooks Paige, James Bell, Aurélien Bellet, Adrià Gascón, Daphne Ezer

▶ **To cite this version:**

Brooks Paige, James Bell, Aurélien Bellet, Adrià Gascón, Daphne Ezer. Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores. Journal of Computational Biology, 2021, 28 (5), pp.435-451. 10.1089/cmb.2020.0445 . hal-03498165

HAL Id: hal-03498165

https://inria.hal.science/hal-03498165

Submitted on 20 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 28, Number 0, 2021
Mary Ann Liebert, Inc.
Pp. 1–17
DOI: 10.1089/cmb.2020.0445

RECOMB 2020

Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores

BROOKS PAIGE,[1,2] JAMES BELL,[1] AURÉLIEN BELLET,[3]
ADRIÀ GASCÓN,[1,4] and DAPHNE EZER[1,4,5]

ABSTRACT

Some organizations such as 23andMe and the UK Biobank have large genomic databases that they re-use for multiple different genome-wide association studies. Even research studies that compile smaller genomic databases often utilize these databases to investigate many related traits. It is common for the study to report a genetic risk score (GRS) model for each trait within the publication. Here, we show that under some circumstances, these GRS models can be used to recover the genetic variants of individuals in these genomic databases—a reconstruction attack. In particular, if two GRS models are trained by using a largely overlapping set of participants, it is often possible to determine the genotype for each of the individuals who were used to train one GRS model, but not the other. We demonstrate this theoretically and experimentally by analyzing the Cornell Dog Genome database. The accuracy of our reconstruction attack depends on how accurately we can estimate the rate of co-occurrence of pairs of single nucleotide polymorphisms within the private database, so if this aggregate information is ever released, it would drastically reduce the security of a private genomic database. Caution should be applied when using the same database for multiple analysis, especially when a small number of individuals are included or excluded from one part of the study.

Keywords: genetic risk scores, genomic privacy, GWAS, long-term privacy, reconstruction attack.

1. INTRODUCTION

In a survey of genomic privacy experts, the long-term privacy of genomic information was deemed both the most important and the most challenging problem to overcome (Mittos et al., 2019). If an individual's password or ID number gets leaked, it is always possible to change it. However, it is impossible for a person to change their genetic code and they will pass part of it onto their children, so any information leaks can have long-term impacts on both the individual and their descendants. Although much of the

[1]The Alan Turing Institute, London, United Kingdom.
[2]Department of Computer Science, University College London, London, United Kingdom.
[3]Inria, Parc Scientifique de la Haute Borne Park Plaza, Villeneuve d'Ascq, France.
[4]University of Warwick, Coventry, United Kingdom.
[5]Department of Biology, University of York, York, United Kingdom.

Brooks Paige, et al., 2021. Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

research focus on long-term privacy of genomic databases rests on the longevity of the encryption scheme (Huang et al., 2015), it is also important to remember that these genomic databases are not just sitting on a server somewhere, but are also being continually utilized for making new scientific discoveries. Each time these databases are accessed and the scientific results are published, there is a risk that information will be leaked and that eventually this would enable an attacker to reconstruct private information held in the database.

Genomic researchers are already aware that some forms of aggregate data from their databases should not be released publicly, because there is a risk that an attacker may be able to determine whether a particular individual is a member of the database (a membership inference attack). For instance, such attacks have already been developed for summary statistics about the frequency of single nucleotide polymorphisms (SNPs; Cai et al., 2015; Dwork et al., 2015; Simmons and Berger, 2015). Membership inference attacks have also been developed for the case where a person is allowed to repeatedly query a database to learn whether at least one individual contains a particular SNP (Shringarpure and Bustamante, 2015; Raisaro et al., 2017; von Thenen et al., 2018). These kinds of aggregate statistics about the frequency or presence/absence of a particular SNP in a database might be useful to release to the broader research community, but it is not an essential output of the research process.

However, the main research findings—that is, the SNPs associated with the trait of interest and their strength of association—are essential to publish since the entire purpose of these genomic research projects is to uncover the relationship between genetic variants and phenotypic traits. Moreover, knowledge of these SNPs can lead to new diagnosis procedures or new potential drug targets, so their release is important for the public interest (Visscher et al., 2017). However, even this information can potentially leak private information about individuals in the database. For instance, Im et al. (2012) found that information about individuals in a genomic database is leaked when studies publish whether each SNP is correlated or anti-correlated to the trait of interest. It is important to quantify how much information is leaked by publishing these research findings, so that scientists can make informed decisions about when to publish their results and whether it is worth risking the privacy of the participants.

In this study, we demonstrate that the kind of research output that is published from genome-wide association studies (GWAS) has the potential to leak enough information to recover the SNPs of individuals in the database (a reconstruction attack), under specific circumstances. In particular, we focus on the release of genetic risk scores (GRS), a common research output for finding genetic associations with continuous traits (Qi et al., 2011; Belsky et al., 2013; Zhao et al., 2014; Chouraki et al., 2016; Day et al., 2017; Knowles and Ashley, 2018). We also focus on cases where a database is repeatedly used to perform a GWAS analysis, but not all the individuals are part of all the analyses. This could be the case because some individuals drop out of the study or skip specific survey questions. Alternatively, some databases, such as 23andMe, may grow in size over time and allow several GWAS to be performed within a short period. Under these circumstances, we demonstrate that it is possible to completely reconstruct the SNPs of an individual by using a custom expectation–maximization (EM) algorithm. We also provide suggestions for avoiding this kind of attack.

To be clear, this article focuses on the simpler case in which the exact same trait is investigated in multiple GWAS studies; however, we expect that some version of this attack may be developed in the near future for the case of multiple highly correlated traits.

1.1. Overview of scenarios that will be investigated

We demonstrate a series of reconstruction attacks that enable us to infer the genotypes of individuals in private genomic databases, based on publicly released GRS. These attacks will initially be deployed in a very favorable scenario, but the scope of the attack will be subsequently expanded, building up to the scenario shown in Figure 1. It is worth noting that the reconstruction attacks that we will describe do not depend on (1) how the SNPs were initially filtered or (2) how strongly they associate with the trait of interest.

We will begin by investigating a simple scenario: Two GWAS studies are performed to identify SNPs associated with the same trait, and the two studies use the same set of participants, except that the second study includes one extra individual. In addition, we will assume that we know the frequencies of each SNP and the frequencies that pairs of SNPs co-occur in the same individual. We assume that both studies publish the coefficients associated with the GRS models that they infer as part of the analysis.

Next, we will consider the case in which the second study includes more than one additional participant and we demonstrate that in many circumstances this still allows us to easily reconstruct the individual genotypes of all the individuals that are found in the second study but not the first (see Section 3.2).

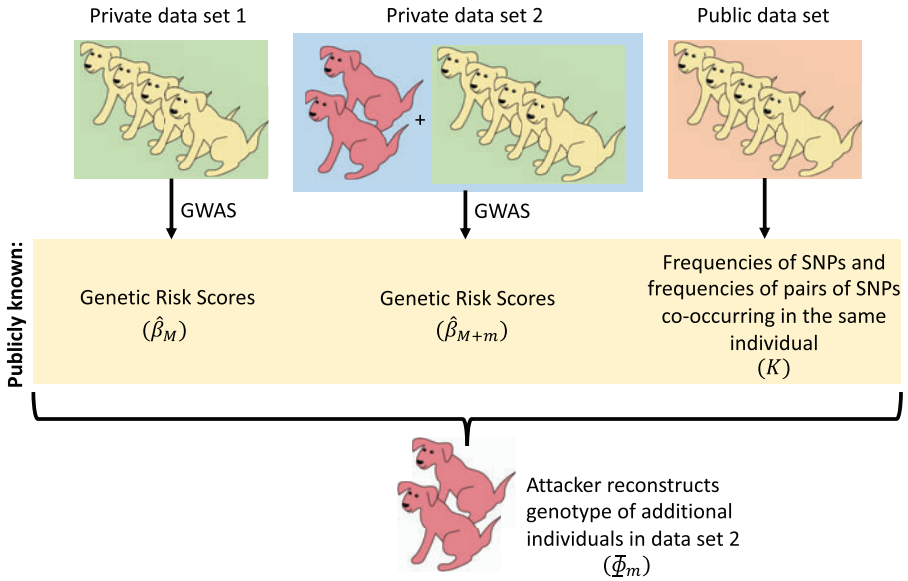


FIG. 1. We investigate the case where two GWAS studies are performed on two datasets that mostly contain the same individuals. We reconstruct the genotype of those individuals added to the second study, using the GRS from each study and an estimate of SNP frequencies. GRS, genetic risk score; GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.

Afterward, we will demonstrate that we do not need to know the precise frequencies of SNPs and frequencies of co-occurring SNPs, as long as we have a reasonable estimate of these values from public databases (see Section 3.3). We also briefly discuss how loosening additional restrictions would impact our ability to predict individual genotypes. In particular, we analyze the case where the two sets of SNPs that are used by the two studies are not identical. These results imply that if two sets of GRS are released on two genetic datasets with largely overlapping populations, it may be possible to reconstruct the genotypes of those individuals who participated in one study but not the other (Fig. 1).

2. METHODS

GRS models describe the relationship between a particular phenotype of interest and particular SNPs. These models are fit in a two-stage process: First, a reduced set of SNPs is selected from a potentially very large pool of candidates; then, this reduced set is used as the independent variables in a linear regression analysis. The set of SNPs is selected by first filtering for those that significantly correlate to the trait of interest, after controlling for other covariates. These SNPs are then further filtered to ensure that they are far apart from one another, to decrease the correlation between them. In this setting, we suppose that M individuals have taken part in a study, and N SNPs have passed the filtration steps to be used in a linear model. Let y_M be the vector of M real-valued phenotypes, and X_M be an $M \times N$ binary matrix, where $X_M[i,j] = 1$ if individual i has SNP j. To include an intercept term in the linear model, we define the design matrix Φ_M to be the $M \times (N+1)$ matrix

$$\Phi_M = [\, X_M \quad 1_M \,]. \tag{1}$$

The GRS model parameter β_M is just the coefficient vector of the linear model

$$y_M = \Phi_M \beta_M + \varepsilon, \tag{2}$$

where ε is independent Gaussian noise. Given Φ_M and phenotypes y_M, the maximum likelihood estimate of this parameter has a closed form

$$\hat{\beta}_M \triangleq \frac{1}{M} K^{-1} (\Phi_M^\top y_M), \tag{3}$$

where we have defined the symmetric $(N+1) \times (N+1)$ matrix K as

$$K = \frac{1}{M} \Phi_M^\top \Phi_M. \tag{4}$$

Now, suppose a second study is run, targeting the same phenotype, which adds a single extra individual with SNPs represented by the N length vector x_0. This corresponds to adding the row $\phi_0^\top = [\, x_0^\top \quad 1\,]$ to the design matrix, and extending y with the additional phenotypic value y_0 for the new individual. The updated estimator (i.e., the GRS values for the second study) is given by

$$\hat{\beta}_{M+1} = (\Phi_M^\top \Phi_M + \phi_0 \phi_0^\top)^{-1} (\Phi_M^\top y_M + y_0 \phi_0). \tag{5}$$

We assume that both GRS models $\hat{\beta}_M$ and $\hat{\beta}_{M+1}$ are released publicly. An attacker aims at using this knowledge to reconstruct ϕ_0 (the genotype of the added individual). Through algebraic re-arrangement (see Section 5.2), we find that:

$$\phi_0 = \frac{1}{C} K(\hat{\beta}_{M+1} - \hat{\beta}_M) \tag{6}$$

where C is a scalar, specifically $C = \frac{1}{M}(y_0 - \phi_0^\top \hat{\beta}_{M+1})$. Eq. (6) means that ϕ_0 is a scalar multiple of $K(\hat{\beta}_{M+1} - \hat{\beta}_M)$.

Our approach, thus, centers on the use of the vector that we define as d_1,

$$d_1 \triangleq K(\hat{\beta}_{M+1} - \hat{\beta}_M) = C\phi_0, \tag{7}$$

corresponding to a rescaled copy of the input SNP data in the design matrix ϕ_0, which can be easily computed from the two parameter vectors if the matrix K is known. As seen in Section 3.1, we can use d_1 to exactly reconstruct the added individual with 100% accuracy.

We additionally consider the case where m additional individuals have been included in the second study, yielding a new GRS model $\hat{\beta}_{M+m}$ including these $M+m$ participants. The extra rows of the design matrix now form a matrix Φ_m of size $m \times (N+1)$, where each row is an individual that was added to the second study and each column is an SNP (and the last column contains only 1). The corresponding analog to Eq. (7) for multiple individuals, which we derive in Section 5.2, is

$$d_m \triangleq K(\hat{\beta}_{M+m} - \hat{\beta}_M) = \Phi_m^\top C_m, \tag{8}$$

where C_m is a vector of length m. For sufficiently small m (relative to N), exact reconstruction of all m added individual genomes is also possible in this setting, following the algorithm we will introduce in Section 3.2.

The previous examples have focused on cases in which the participants in the first study are a subset of the individuals in the second study. In Section 5.7, we consider the case in which the first study has some participants that are not found in the second study and vice versa. We show that the same strategies for reconstructing the genome can be used as in the previous scenario that we discussed, in which multiple participants are added to the second study.

2.1. Estimation of K

As it turns out, the entries of matrix K correspond to simple population-level statistics of the SNPs, which could either be inadvertently released (under the assumption they would be safe to share), or could be estimated from another sample from the same population. In fact, the entries of K depend only on the SNP frequencies and SNP co-occurrence frequencies in the dataset:

– For $i = 1, \ldots, N$: K_{ii} estimates the probability that SNP i has value 1 (i.e., the frequency of the SNP in the population).
– For $i = 1, \ldots, N-1$ and $j > i$: $K_{ij} = K_{ji}$ estimates the probability that both SNP i and SNP j are 1 simultaneously (i.e., the frequency of SNP i and SNP j co-occurring in the same individual).
– For $i = 1, \ldots, N$ and $j = N+1$: $K_{ij} = K_{ji}$ also estimates the probability that SNP i has value 1, that is, $K_{i,N+1} = K_{N+1,i} = K_{ii}$.
– Finally, $K_{N+1,N+1} = 1$.

Thus, knowledge of SNP frequencies and pairwise co-frequencies from the original study are all that is required to compute K. In the following Sections 3.1 and 3.2, we consider adding one and multiple individuals at once, respectively, in the setting where this matrix K can be estimated exactly. However, although $\hat{\beta}_M$, $\hat{\beta}_{M+1}$ and M are likely to be published along with the study, an attacker would often need to estimate K from other publicly available data. Most studies will report some information about the study population (such as whether the study focused on individuals from a specific continent), which can help with estimating K. From this information, we can estimate the value of K in similar populations as those used in the study using publicly available data, for example, from the HapMap project. Our additional experiments in Section 3.3 use a custom EM algorithm to find maximum likelihood estimates of ϕ_0 when the matrix $\hat{K} \approx K$ is estimated from independent public data. The derivation of this EM algorithm is given in Section 5.4.3, and a formal analysis of the reconstruction error of ϕ_0 given the error in $\hat{K}$ is found in Section 5.2.1.

3. RESULTS

The key observation from the previous section is that the vectors d_1 and d_m, derived from the change in parameter vectors $\hat{\beta}$ from a first study to a second study, *take only a finite number of values* thanks to the fact that the design matrices Φ contain only zeros and ones. In particular, when m new individuals are added to the second study, each entry of the vector d_m can only take at most 2^m values, and a zero value corresponds to the setting where all individuals have the most common variant for that SNP.

This section describes algorithmically how these vectors can be used to recover the genomes of the additional individuals, as well as empirical tests that use the Cornell Dog Genome dataset as a case study (Hayward et al., 2016). More details on the experimental setup can be found in Section 5.1.

3.1. Complete reconstruction of one individual's genotype when SNP frequency information is known

The first, most straightforward case is when only one participant is added between the first and second studies, that is, where $\hat{\beta}_M$ is the GRS for the first study (containing M participants), and $\hat{\beta}_{M+1}$ is the GRS for the second study as described in Eqs. (3) and (5). Both of these are vectors of length $N+1$, where the first N indices correspond to the relationship between each SNP and the trait and the last element is the intercept of the linear model. For now, we also assume we are in the setting where the matrix K is known, for example, because the SNP frequency information has been publicly released.

Given K, $\hat{\beta}_{M+1}$, and $\hat{\beta}_M$, we can use d_1 (a vector of length $N+1$) to precisely determine the genotype of the individual who was added to the database. For each $i = 1, \ldots, N$, the i^{th} entry of d_1 is either equal to 0 if ϕ_0 contains a 0 (i.e., the individual does not have the SNP at that index) or to C if ϕ_0 contains a 1 (i.e., the individual has the SNP at that index). In other words, it is possible to exactly read off the SNPs of the added individual in this setting. Indeed, we tested this strategy on the Cornell Dog Database and found that we were able to reconstruct the genotype of the dog that was added to the second study with 100% accuracy, on both common and uncommon SNPs (Fig. 2A).

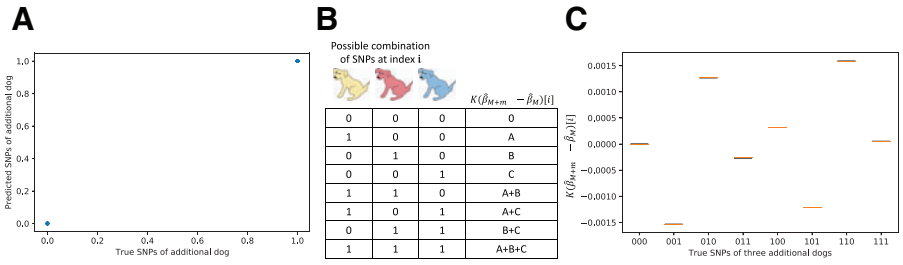


FIG. 2. **(A)** We have perfect accuracy in reconstructing the genotype when K is known (using 200 random SNPs to estimate average breed weight in the Cornell Dog Database). **(B)** We can reconstruct all the genotypes of multiple dogs that are added to the second study and **(C)** this works in practice by using the data from the Cornell Dog Database, as in **(A)**.

3.2. Complete reconstruction of multiple individuals' genotype when SNP frequency information is known

We now consider the case where m additional individuals have been included in the second study, yielding a new GRS model $\hat{\beta}_{M+m}$, including these $M+m$ participants.

Consider again Eq. (8) described earlier. The i^{th} row of Φ_m is a binary vector that represents the combination of the m individuals who have SNP i. This means that, for a fixed value of C_m, the value of the vector d_m at index i is uniquely determined by the combination of individuals who have SNP i (Fig. 2B). In other words, there will be at most 2^m unique values taken by entries of d_m, each corresponding to a combination of the values in vector C_m (Fig. 2C).

If we were to learn which values of d_m are also found in C_m, then we could infer the complete genotypes of all the m individuals added to the second study. We would be able to reconstruct m complete genotype vectors, although it would be impossible to know which of the genotypes corresponded to which of the m individuals. In fact, in many cases it is extremely straightforward to determine which values in d_m correspond to values in C_m. Here, we describe a simple algorithm for finding C_m when there are exactly 2^m unique values in d_m. If this is not the case, please refer to the more complete algorithm in Section 5.3.

1. First, extract all unique, nonzero values from d_m.
2. Find the sum of all pairs of values in this vector.
3. Find all values that are in Eq. (1), but not in Eq. (2). The values of C_m appear in this list. There is no way to know which value of C_m corresponds to which index, so for simplicity we can randomly assign them indices.
4. Each value in C_m corresponds to a specific individual who was added to the second study. Each value in d_m can be described as a sum of a unique combination of values in C_m. For instance, if $d_m[i] = C_m[j] + C_m[k]$, this means that the SNP at position i is found in individual j and k, but no one else.

We tested this approach by using the Cornell Dog Database, in a test scenario where the second study added three different dogs. We were able to uniquely identify the genotypes of all three dogs with 100% accuracy, with both common and uncommon SNPs (Fig. 2C).

3.3. Accurate estimation of an individual's genotype when SNP frequency information is estimated from a public database

Previously, we assumed that the attacker had access to the matrix K, which consists of population-level statistics on frequencies and co-occurrence frequencies of SNPs. Although this could be released voluntarily by organizations that are not aware of the risk, we now consider the case where K is not directly available to the attacker but is instead estimated from a separate public database assumed to correspond to individuals from the same population.

We simulated this scenario by using the Cornell Dog Database by taking one random set of dogs for building the GRS model, and a second non-overlapping set of dogs for estimating $\hat{K}$. We compared the value of $\hat{d}_1 = \hat{K}(\hat{\beta}_{M+1} - \hat{\beta}_M)$ with the known value of ϕ_0. We observe that $\hat{d}_1$ has significantly different values at indices where $\phi_0[i] = 0$ and $\phi_0[i] = 1$; examples for the cases where one and three dogs are added can be seen in Figure 3.

The main challenge is that the vector $\hat{d}_1$ now includes additional noise, so we cannot simply use its entry at index $N+1$ to estimate C, nor do the entries i with $\phi_0[i] = 0$ also correspond directly to $\hat{d}_1[i] = 0$. Instead, we develop a custom EM algorithm to find a maximum likelihood estimate of the constant C and recover ϕ_0, that is, to determine the probability that each $\phi_0[i] = 0$ or $\phi_0[i] = 1$, based on the value of $\hat{d}_1$ (see Section 5.4.3 for details). We find that this method can successfully reconstruct the correct value of $\phi_0[i]$ much better than a baseline that uses the public dataset to independently estimate the most common variant for each of the SNP (Fig. 4).

Crucially, we show that our approach is able to reconstruct, with relatively high accuracy, the genotypes of dogs even when they differ significantly from those in the public dataset (Fig. 5). This shows that our attack is able to extract information about the particular individuals that differ across the two studies, not merely about the general population as in the most-common-variant baseline. By definition, dogs that have genotypes that differ significantly from the general population have a higher proportion of uncommon SNPs, and the ability to recover these uncommon SNPs is particularly important from a privacy perspective. Indeed, uncommon SNPs can be used to identify a particular individual and are also more likely to be associated with disease phenotypes, which is sensitive information. In general, we find that the larger the

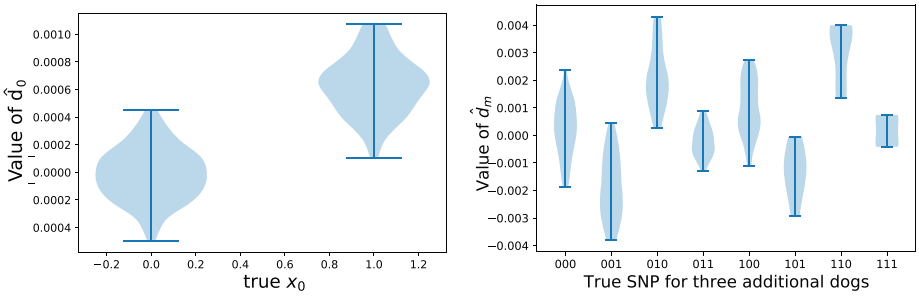


FIG. 3. Example values taken by the noisy vector $\hat{d}$, given the true value of the corresponding SNP in the genome. (Left) adding one new participant; (right) adding three new participants. These figures are analogous to those in Figure 2, although in the case where K is not known and instead estimated from an independent public database.

public dataset available, and the more similar the dataset is to the unknown private dataset, the better we are able to reconstruct the genome of the added individual. Full details and description of the experimental setting are given in Section 5.1. We also derive theoretical error bounds for our estimate of ϕ_0 based on the error in $\hat{K}$ in Section 5.4.1.

This task becomes more challenging when multiple individuals are added simultaneously and K is unknown; an algorithm for estimating Φ_m for $m > 1$, along with additional empirical results, is given in Section 5.4.

3.4. Accurate estimation of an individuals' genotype when different SNPs are used in each study

When GRS models are constructed, the first step is to filter the set of SNPs down to a small set of SNPs that are (1) significantly correlated to the trait after covariates are considered and (2) far apart from one another along the genome. If the two studies use two different sets of SNPs to construct the GRS model, it is still possible to recover whether or not each of the SNPs in the overlap is present in the new individual. This process is highly analogous to the previous cases and is detailed in Section 5.6.

4. DISCUSSION

In this study, we demonstrate that private information is leaked when GRS models are published, specifically in the case where two sets of largely overlapping individuals are used for multiple studies. In particular, we show that we can recover SNPs from an individual in a private database—a reconstruction attack. Even though we would not have a name associated with this genotype, it may be possible to identify the individual once the genotypic data are available to the attacker. For instance, the attacker may have

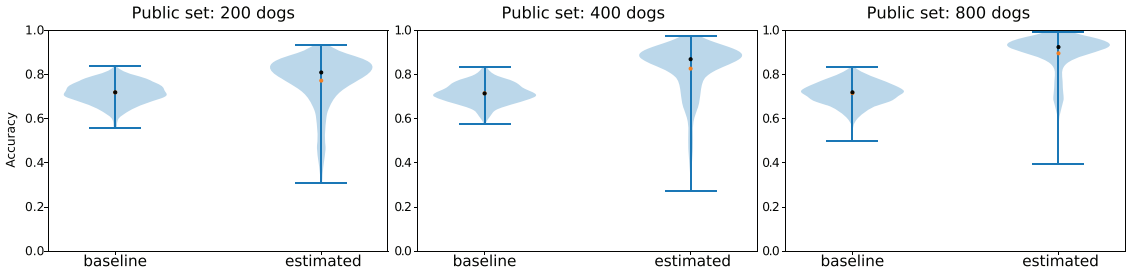


FIG. 4. Accuracy at reconstruction of genomes x_0 using EM estimation and a noisy estimate $\hat{K}$, as compared with a natural baseline that always predicts the most common variant at each SNP locus. We use this as a baseline, because without any additional information about β_M and β_{M+1}, the most accurate prediction of the dog's genotype would be to predict the most common variant at each locus. Here, we define accuracy as the proportion of SNPs that are correctly identified in the dog that was found in the second GWAS study, but not the first. Each distribution is constructed from 500 experimental test points, in which we (1) took 10 random splits of the full dog dataset, assigning dogs to either the public or private dataset; (2) for each split, we tested the reconstruction 50 times, each time adding a different randomly sampled dog to the second GWAS study. The private dataset always has 1000 individuals; the public test dataset is of increasing size, improving performance. EM, expectation–maximization.

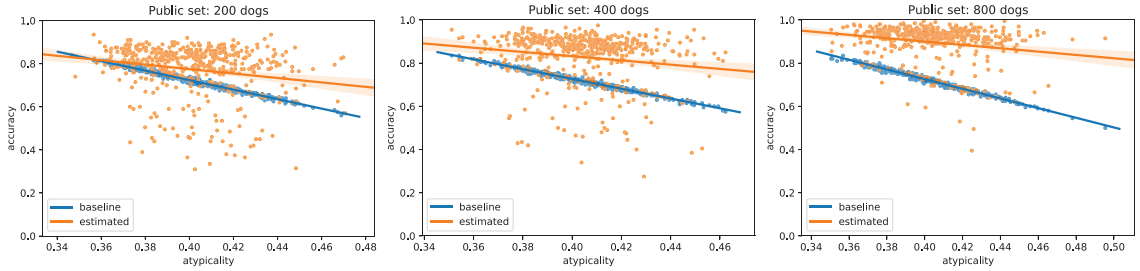


FIG. 5. Results of Figure 4 broken down by individual dogs. Here, each point represents a dog and we define *atypicality* as the proportion of *uncommon variants* that the dog has compared with the public database—for instance, if 51% or more of dogs in the public database have a G in a specific locus, but this dog has a T, then this would count toward the dog's atypicality. In other words, dogs further to the right are less and less similar to average dogs present in the public dataset (measured by percentage of different variants). In contrast to the most-common-variant baseline, our method generalizes well even to dogs that are highly dissimilar to those in the public dataset. Larger public databases (right) provide more accurate population estimates of $\hat{K}$, leading to more accurate reconstructions overall.

access to partial genotypic information of the individual and then be able to identify them. Alternatively, they could use the genotype information to predict ethnicity and other phenotypic traits that could then be used to uniquely identify the individual.

We also note that even an incomplete reconstruction attack (in which only a proportion of the SNPs are correctly identified) is likely to be sufficient to perform a membership inference attack. Investigating the relationship between the reconstruction attack and the membership inference attack will be a subject of future research. Importantly, if the attackers were unable to link the genomic data with a particular individual, the reconstruction attack would still be a breach in privacy that could have serious consequences. For instance, the patient may have only consented to have their genomic data used in particular kinds of research studies, whereas the attacker may use the reconstructed genomic data for a different (potentially unethical) purpose.

4.1. Suggestions for good practice

We provide a number of simple suggestions for good practice that would help limit this attack.

1. Aggregate statistics about the frequency of SNPs in the database or the frequency of co-occurrence of SNPs should never be released. We have shown that this information, combined with GRS, allows to precisely reconstruct individual genomes in various settings. It may be possible to release *noisy* versions of SNP frequency data, but this would be equivalent to releasing $\hat{K}$ (our estimated K from the public database). With our EM algorithm, we have demonstrated that it is still possible to do some genotypic reconstruction with a noisy $\hat{K}$, but this becomes harder as the noise in $\hat{K}$ increases. On the other hand, providing a very noisy $\hat{K}$ may be of limited utility to the scientific community.
2. If a genetic dataset is intended to serve for multiple complementary analyses, it is important that all study participants are used in every analysis performed. If there are missing phenotypic data from a few individuals, they should not be included in *any* of the analyses performed, or their privacy may be compromised.
3. When multiple individuals are added in between two studies, then the ability to reconstruct the genomes depends on the number of SNPs, being large relative to the number of individuals. In particular, if m new dogs are added, exact reconstruction is only possible by using the approach in Section 3.2 if the number of SNPs $N > 2^m$. Thus, we suggest to avoid releasing multiple studies that differ by fewer than $\log_2 N$ individuals.

4.2. Extensions and future work

Although we have analyzed the case where the genome is represented by binary values of 0 or 1, often studies instead count the number of times each allele is present, which would lead to a design matrix Φ containing values 0, 1, or 2. In this scenario, K no longer contains the frequencies of SNPs and their co-occurrences, but it is something slightly more complicated that we describe in Section 5.8. This does not dramatically change the approach in this study, except in that the vector d_m can take 3^m possible values,

rather than 2^m. In practice, then, studies that use allele counts are somewhat more robust to attacks; the multiple dog reconstruction attack would likely be ambiguous if $3^m > N$, rather than $2^m > N$.

A possible countermeasure to our reconstruction attack could consist of randomly perturbing the GRS models before releasing them, as done in differentially private linear regression (Wang, 2018). However, a naive application of this strategy could destroy the utility of the models. A formal and empirical analysis of the effectiveness of such protection against reconstruction attacks, as well as of the usefulness of the resulting GRS models to genomic researchers, is beyond the scope of this article and left for future work.

Another countermeasure is to refrain from releasing precise information about the population structure of the study population to prevent the attacker from estimating K effectively. This would, however, limit the utility of the research study, because the researchers would not know to what populations the research applies.

Our work has a number of limitations. For instance, we only test our EM algorithm on dog data. Dog populations may have different population structures than human populations due to selective breeding, so in the future we aim at investigating how properties of population structure will impact our ability to estimate K and the accuracy of our reconstruction attack.

It may seem unlikely on the surface that two GWAS analyses will include nearly the same participants. One potentially common setting where this could arise is when a single study collects both genotype and phenotype data from a single set of participants, and it releases multiple models to predict multiple traits. In this case, there may be a small number of individuals who are used in one analysis, but not the other; for instance, there may be a small subset of participants who skip a particular survey question that was used to collect phenotype information, and this is, indeed, evident in a recent study (Jiang et al., 2019). In such settings, it could be very possible for multiple released GRS models to be computed on sets of individuals that differ by only a few participants. In future work, we aim at extending our analysis and attack to settings where multiple GRS models are released, each predicting different but highly correlated traits.

5. APPENDIX

5.1. Experimental details

5.1.1. Cornell dog database. To experimentally test the reconstruction attacks, we used data from the Cornell Dog Genome Database, which contains data about SNPs from a wide range of dog breeds and a number of associated phenotypic traits. The two traits we focused on were *average breed weight* and *average breed height*, because these two phenotypes had the fewest number of missing values. For the initial investigation, we binarized the genotype matrix—considering all heterogenous alleles to have a value of 1. (We also repeated the analysis with the original genotype matrix.) Only common SNPs (i.e., SNPs that were found in 25%–75% of the dogs) were used, leaving 23,497 SNPs. For each linear model built, $M = 1000$ dogs were randomly sampled as the ''private'' dataset and $N = 200$ SNPs were randomly selected. To ensure that the SNPs that were sampled were spatially distributed, the SNPs were randomly sampled in a stratified way, so one SNP was selected in every $\frac{23,497}{200}$-sized bin.

5.1.2. Experiment with imprecise K. First, two linear models were constructed to predict average breed weights: one with the $M = 1000$ randomly sampled dogs and another that contained one additional randomly sampled dog. This gives $\hat{\beta}_M$ and $\hat{\beta}_{M+1}$. To mimic the process of estimating K from a public database, we randomly sampled an additional 200, 400, or 800 dogs that were not included as part of the original set and used this to estimate K, which we denote by $\hat{K}$. Now, we could calculate $\hat{K}(\hat{\beta}_{M+1} - \hat{\beta}_M)$ and compare this with the known ϕ_0 for the additional dog from the second study. These additional dogs are taken from a third ''test'' dataset, disjoint from both the public and private data. The plots in Figures 4 and 5 are produced by re-running the algorithms across 10 random public/private/test splits, where the ''test'' dataset has 50 dogs that are each individually considered as candidates for the $(M+1)^{\text{th}}$ dog added to the private dataset.

5.2. Adding multiple dogs

Here, we explain Eqs. (7) and (8). Note that the former is a special case of the latter so we will only explain the latter in detail. First note that, by definition,

$$\hat{\beta}_M = (\Phi_M^\top \Phi_M)^{-1} \Phi_M^\top y_M = (MK_M)^{-1} \Phi_M^T y_M,$$
$$\hat{\beta}_{M+m} = (\Phi_{M+m}^\top \Phi_{M+m})^{-1} \Phi_{M+m}^\top y_{M+m} = (MK_M + mK_m)^{-1} \Phi_{M+m}^\top y_{M+m}.$$

Substituting these into the left hand side of the following equation gives the right hand side:

$$(MK_M + mK_m)\hat{\beta}_{M+m} - MK_M\hat{\beta}_M = \Phi_m^T y_m. \quad (9)$$

This equation can be rearranged to give

$$\begin{aligned} K_M(\hat{\beta}_{M+m} - \hat{\beta}_M) &= \frac{1}{M}\Phi_m^T y_m - \frac{m}{M}K_m\hat{\beta}_{M+m} \\ &= \frac{1}{M}\Phi_m^T(y_m - \Phi_m\hat{\beta}_{M+m}). \end{aligned}$$

Defining the length m vector $C_m = \frac{1}{M}(y_m - \Phi_m\hat{\beta}_{M+m})$ yields the form used in Eq. (8). For the special case of $m=1$, C_m is a scalar and we recover Eq. (7).

5.3. *Algorithm for identifying unique genotypes of multiple dogs when* K *is known*

Although the simple approach described in the main article will work in many cases, there are a few special circumstances where a more complex algorithm may be required. In particular, it would not work if there are combinations of SNPs that are not observed among the individuals added to the database. For instance, if there is not an SNP location where the first individual has an SNP variant and the others do not, then we would miss the corresponding value in C_m. However, it is still possible to identify all the values in C_m through a more complex algorithm:

1. First, extract all unique, nonzero values from d_m.
2. Find the sum of all pairs of values in Eq. (1).
3. Find all values that are in Eq. (1), but not in Eq. (2).
4. If there are exactly m values in Eq. (3) and the sum of these values equal the last value of d_m (corresponding to the intercept term), then we have found the correct values of C_m.
5. Otherwise, this suggests that there are one or more elements of C_m that are missing from Eq. (3) and possibly a few values in Eq. (3) that are not in C_m.
6. Begin by subtracting every pair of values in Eq. (3). These are now also potential values of C_m
7. Search for a set of m values from Eqs. (3) and (6) that sum to the last element of d_m. There may be more than one set of values for which this is true.
8. If this search is unsuccessful, repeat steps 6–7. Eventually, a set of m values summing to d_m should be found.
9. If more than one possible set of values is found for C_m in Eq. (7), it is still possible to compare these sets and identify which is the most likely to contain the true values of C_m. For each possible C_m vector, a set of genotypes can be constructed for the m additional individuals. Using the frequencies of each SNP, it is possible to calculate the probability of observing each genotype. The set of values that produces the most likely genotypes for the m individuals is most likely to be the correct one.

In addition, this algorithm depends on the fact that it is extremely unlikely that if someone were to sample three random continuous numbers i, j, and k, it would just so happen that $i+j=k$. There is an extremely small chance that a value of C_m would be un-discoverable because of a coincidence of this nature.

5.4. *Estimating* K

If the true matrix K is unknown, it can be estimated with public data. We denote this estimator by $\hat{K}$. For $\hat{K}$ to be an accurate estimate, the data that it is generated from must be drawn from the same (or a sufficiently similar) population as that used in the private study. We will model this assuming no discrepancy between population distributions; however, when we discuss how to evaluate whether the estimate is good, that assessment should account for this systematic error as well. Later, we will be primarily concerned with the error due to the subsampling in both the private and public datasets.

Also of note, the same analysis given next applies to the scenario in which the researchers do not release K, but rather release a ''noisy'' version of K, where the noise is drawn from a normal distribution. They might consider doing this if they feel that releasing information about SNP frequencies is important for the research community, but they do not wish to release the real K because this would allow for an exact reconstruction of genotype. This noisy K could still be used in a reconstruction attack in the same way as an estimate of K from a public database is used.

5.4.1. Analytic bound on $\| \phi_0 - \hat{\phi}_0 \|$. For convenience, we only consider the case of adding a single individual, though the generalization is quite straightforward. If $\hat{K}$ is substituted for K in our reconstruction Eq. (7), we get an approximation of ϕ_0 that we denote $\hat{\phi}_0$. We would like to bound the (relative) error between ϕ_0 and $\hat{\phi}_0$. Later, we ignore the constant factors C and $\hat{C}$ for simplicity, noting that these scaling factors are estimated from the resulting ϕ_0 or $\hat{\phi}_0$ anyway. We, thus, consider $\varphi_0 = K(\hat{\beta}_{M+1} - \hat{\beta}_M)$ and $\hat{\varphi}_0 = \hat{K}(\hat{\beta}_{M+1} - \hat{\beta}_M)$. Using $\| \cdot \|$ on vectors, and also on matrices, we denote the corresponding operator norm. The relative error between φ_0 and $\hat{\varphi}_0$ is given by:

$$\frac{\| \varphi_0 - \hat{\varphi}_0 \|}{\| \varphi_0 \|} = \frac{\| (\hat{K}\hat{K}^{-1} - \hat{K}K^{-1})\varphi_0 \|}{\| \varphi_0 \|} \leq \| \hat{K}\hat{K}^{-1} - \hat{K}K^{-1} \| = \| \hat{K}(\hat{K}^{-1} - K^{-1}) \| .$$

Note that $\hat{K}^{-1} - K^{-1} = \hat{K}^{-1}(K - \hat{K})K^{-1}$ and hence

$$\frac{\| \hat{\varphi}_0 - \varphi_0 \|}{\| \varphi_0 \|} \leq \| K^{-1} \| \| K - \hat{K} \| , \tag{10}$$

This means that we can bound the error by two quantities. The term $\| K^{-1} \|$ is bounded earlier by $1/\min(eig(K))$, which is finite as soon as K is nonsingular. This is not a strong requirement, as in the case of linear regression it is required for $\hat{\beta}_{M+1}$ and $\hat{\beta}_M$ to exist. Note that in the case of L2-regularized linear regression (i.e., ridge regression), K is replaced by $K + \lambda I$, where λ is the regularization parameter, and we can directly bound this term by λ.

The key term in Eq. (10) is $\| K - \hat{K} \|$, the error in estimating K by $\hat{K}$. Let us assume that the public database used to obtain $\hat{K}$ follows the same distribution as the private database used to fit the GRS models. Denote by $\hat{M}$ the number of individuals used to estimate $\hat{K}$. Then, under classic boundedness assumptions and leveraging matrix concentration inequalities such as matrix Bernstein Tropp (2015), we can show that $\mathbb{E}[\| K - \hat{K} \|] = O(1/\sqrt{\min(\hat{M}, M)})$. This shows that the error in estimating K is small as long as the private and public databases are large enough.

5.4.2. Modeling the error in $\hat{K}$. In this section, we define a model to capture the error in $\hat{K}$, which leads to the EM algorithm for estimating ϕ_0, which is used in the experiments. As our estimated $\hat{K}$ drifts from the true K, this expression $\hat{K}(\hat{\beta}_{M+1} - \beta_M)$ would produce a wider range of values than just 0 and C.

Let $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ be independent noise, which we assume corrupts each element of K_{ij}; that is, given the estimated matrix $\hat{K}$, suppose

$$K_{ij} \sim \mathcal{N}(\hat{K}_{ij}, \sigma^2), \tag{11}$$

for some small σ^2. This is clearly an oversimplification (as we know K is e.g., bounded and symmetric), but it is a useful starting point that allows derivation of a simple estimation algorithm. For notational brevity, in this and the following section we define the vector

$$\Delta = \hat{\beta}_{M+m} - \hat{\beta}_M \tag{12}$$

which corresponds to the difference between the two GRS model parameter vectors when m additional dogs are added. Given the true value of K, the system of equations

$$\Phi_m^\top c_m = K\Delta$$

relates the known quantity Δ and the Gaussian-distributed K with the matrix Φ_m and the unknown values in the vector $c_m \in \mathbb{R}^m$. This breaks down into a sum across the entries in c_m, with

$$K\Delta = \sum_{j=1}^{m} C_j \phi_j .$$

We need to estimate all m constants $C_j, j = 1, \ldots, m$.

If K is Gaussian [following Eq. (11)], then the linear transformation $K\Delta$ is Gaussian as well. We denote each of the rows of K as a vector k_i, $i = 1, \ldots, N$; then, for each row, the scalar value

$$k_i^\top \Delta \sim \mathcal{N}(\hat{k}_i^\top \Delta, \sigma^2 \Delta^\top \Delta),$$

meaning overall the vector $K\Delta$ is distributed $\mathcal{N}(\hat{K}\Delta, \sigma^2 \Delta^\top \Delta I)$.

With some algebraic re-arrangement, and since for the true underlying value of K we have $K\Delta = \Phi_m^\top c_m$, we can write this as

$$\hat{K}\Delta \sim \mathcal{N}(\sum_{j=1}^{m} C_j \phi_j, \sigma^2 \Delta^\top \Delta I) \tag{13}$$

where $C_1, \ldots, C_m$ and σ are parameters we need to estimate. The vector $\hat{K}\Delta$ is observed "data," computed from the public SNP database and the two released parameter vectors. We can model each of the entries of Φ_m, which are zeros and ones, as Bernoulli distributions, whose prior probabilities correspond to the public dataset estimated frequencies. This suggests a model for $\hat{K}\Delta$ that is akin to a constrained mixture of Gaussians.

For the special case of $m = 1$, with only a single scalar C and vector ϕ_0, this reduces to

$$\hat{K}\Delta \sim \mathcal{N}(C\phi_0, \sigma^2 \Delta^\top \Delta I). \tag{14}$$

5.4.3. Parameter estimation with EM. We now can use this model to derive EM algorithms for finding maximum likelihood estimates of all parameters, and estimate the posterior distribution over SNP variants for the added individuals between the two studies.

For notational convenience in this section, denote the entries of the m new individuals $\Phi_m \in \{0, 1\}^{N+1, m}$ as $z_{i,j}$, for $i = 1, \ldots, N+1$ and $j = 1, \ldots, m$, and let z_j denote the column vector $z_{1,j}, \ldots, z_{N+1,j}$. Denote the prior probabilities for each i as $\alpha_1, \ldots, \alpha_{N+1}$, where $\alpha_1, \ldots, \alpha_N$ are the (public) population frequencies for each SNP, and $\alpha_{N+1} = 1$. Let $x_1, \ldots, x_{N+1}$ denote the entries of the fixed (observed) vector $x = \hat{K}\Delta$, which in this simplified notation is distributed as

$$p(x|c_m, \Phi_m, \sigma^2) = \mathcal{N}(x|\sum_{j=1}^{m} C_j z_j, \sigma^2 \Delta^\top \Delta I).$$

Supposing we know values of $C_1, \ldots, C_m, \sigma^2$, to estimate the entries of Φ_m we want to find $p(z|x, C, \sigma^2)$,

$$p(z|x, c_m, \sigma^2) \propto p(x|c_m, \Phi_m, \sigma^2) p(z).$$

An EM algorithm to estimate c_m, σ^2 would proceed by alternately:

1. Given c_m, σ^2, estimate the posterior distribution $\pi = p(z|x, c_m, \sigma^2)$;
2. Given the posterior π, maximize $\mathcal{L} = E_\pi[\log p(x|c_m, \Phi_m, \sigma^2)]$ with respect to c_m and σ^2.

For each $z_{i,j}$, we can analytically compute the distribution

$$\begin{aligned} &p(z_{i,j} = 1|x, c_m, z_{k \neq j}, \sigma^2) \\ &= \frac{\alpha_i \mathcal{N}(x_i|C_j + \sum_{k \neq j} C_k z_{i,k}, \sigma^2 \Delta^\top \Delta)}{\alpha_i \mathcal{N}(x_i|C_j + \sum_{k \neq j} C_k z_{i,k}, \sigma^2 \Delta^\top \Delta) + (1 - \alpha_i) \mathcal{N}(x_i|\sum_{k \neq j} C_k z_{i,k}, \sigma^2 \Delta^\top \Delta)}, \end{aligned} \tag{15}$$

the conditional probability of each particular entry taking a value of 1, rather than 0, for each z_j given the values of the other z_k, $k \neq j$. Note that each SNP location i can be treated independently; however, each of the individuals $j = 1, \ldots, m$ individuals must be considered jointly.

5.4.4. Exact EM algorithm when 1 individual is added. For the special case of $m = 1$, this yields a tractable exact EM algorithm. Since there are no other individuals, Eq. (15) reduces to $p(z|x, C, \sigma^2)$, with

$$\pi_i = p(z_i = 1|x, C, \sigma^2) = \frac{\alpha_i \mathcal{N}(x_i|C, \sigma^2 \Delta^\top \Delta)}{\alpha_i \mathcal{N}(x_i|C, \sigma^2 \Delta^\top \Delta) + (1 - \alpha_i) \mathcal{N}(x_i|0, \sigma^2 \Delta^\top \Delta)} \tag{16}$$

the posterior probability of each particular entry taking a value of 1, rather than 0. To maximize $\mathcal{L} = E_\pi[\log p(x|C, \phi_0, \sigma^2)]$ with respect to C and σ^2, we first compute the derivatives of

$$\begin{aligned} \mathcal{L} &= \sum_i \sum_{z_i} p(z_i| \ldots) \log p(x_i|C, z_i, \sigma^2) \\ &= \sum_{i=1}^{N} \pi_i \log \mathcal{N}(x_i|C, \sigma^2 \Delta^\top \Delta) + (1 - \pi_i) \log \mathcal{N}(x_i|0, \sigma^2 \Delta^\top \Delta). \end{aligned}$$

This yields

$$\frac{\partial \mathcal{L}}{\partial C} = \sum_i \frac{\pi_i}{\sigma^2 \Delta^{\mathrm{T}} \Delta}(x_i - C),$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \sum_{i=1}^{N} \pi_i \frac{\partial}{\partial \sigma^2} \log \mathcal{N}(x_i | C, \sigma^2 \Delta^\top \Delta) + (1-\pi_i) \frac{\partial}{\partial \sigma^2} \log \mathcal{N}(x_i | 0, \sigma^2 \Delta^\top \Delta),$$

which we set equal to zero and solve to find

$$\hat{C} = \frac{\sum_i \pi_i x_i}{\sum_i \pi_i}, \tag{17}$$

$$\hat{\sigma}^2 = \frac{1}{N \Delta^\top \Delta} \sum_{i=1}^{N} \pi_i (x_i - C)^2 + (1-\pi_i) x_i^2. \tag{18}$$

These updates taken together can be used to define an EM algorithm that optimizes the values of C and σ^2, despite the fact that the entries of ϕ_0 are unknown; once C and σ^2 are then known, the vector π will give probability estimates for each entry of ϕ_0.

The overall EM algorithm can be summarized by the following iterative updates:

1. $\pi_i \equiv p(z_i = 1 | x, \hat{C}, \hat{\sigma}^2) = \frac{\alpha_i \mathcal{N}(x_i | \hat{C}, \hat{\sigma}^2 \Delta^\top \Delta)}{\alpha_i \mathcal{N}(x_i | \hat{C}, \hat{\sigma}^2 \Delta^\top \Delta) + (1-\alpha_i) \mathcal{N}(x_i | 0, \hat{\sigma}^2 \Delta^\top \Delta)}$,
2. $\hat{C} \leftarrow \frac{\sum_i \pi_i x_i}{\sum_i \pi_i}$,
3. $\hat{\sigma}^2 \leftarrow \frac{1}{N \Delta^\top \Delta} \sum_{i=1}^{N} \pi_i (x_i - \hat{C})^2 + (1-\pi_i) x_i^2$.

To initialize the algorithm, we can set π_i to some initial probabilities, and find initial values for $\hat{C}$, $\hat{\sigma}^2$; we experimented with setting both to the prior probabilities per SNP estimated from the public data and to the vector of all zeros (corresponding to a ''hard'' initialization at the value of the baseline estimate), and we found no qualitative difference in performance.

5.4.5. Stochastic EM (SEM) for multiple individuals. For $m > 1$, the exact posterior depends on all individuals and does not have a compact form. However, we can easily approximate the posterior by Gibbs sampling using Eq. (15), which describes the full conditional distribution $p(z_{i,j} = 1 | x, c_m, z_{k \neq j}, \sigma^2)$, iteratively drawing samples for each individual j. We can use this for parameter estimation of σ^2 and each $C_1, \ldots, C_m$ by using the stochastic EM algorithm Celeux and Diebolt (1985), which differs from a standard EM algorithm in that the expectation step (evaluating the posterior) is replaced by Monte Carlo sampling. In this algorithm, we alternately

1. draw approximate posterior samples of $z_{i,j}$ by one or more sweeps of Gibbs sampling, following Eq. (15);
2. conditioned on the current sampled values $z_{i,j}$, find values of σ^2 and $C_1, \ldots, C_m$ which maximize the likelihood $\mathcal{N}(x_i | \sum_{j=1}^{m} C_j z_{i,j}, \sigma^2 \Delta^\top \Delta I)$.

Although this does not converge to an exact parameter value, under suitable conditions the algorithm converges in distribution to a Gaussian centered on the maximum likelihood estimate of the parameter. A point estimate can be extracted by averaging across many iterations after convergence.

In contrast to the EM updates, the updates for values of C_j and σ^2 given actual sampled values of z_j are straightforward and do not scale combinatorially in m. Optimizing c_m corresponds to solving a least-squares problem, that is,

$$\min_{C_1, \ldots, C_m} \sum_{i=1}^{N+1} (x_i - \sum_{j=1} C_j z_{i,j})^2 = \min_{c_m} \| x - Z c_m \|_2^2 ,$$

using the vector notation $c_m = [C_1, \ldots, C_m]^\top \in \mathbb{R}^m$, $x = [x_1, \ldots, x_{N+1}]^\top \in \mathbb{R}^{N+1}$, and $Z \in [0, 1]^{N+1, m}$, has the solution

$$\hat{c}_m = (Z^\top Z)^{-1} Z^\top x. \tag{19}$$

The maximum likelihood estimate of σ^2 given this estimated $\hat{c}_m$ is simply the mean squared error

$$\hat{\sigma}^2 = \frac{1}{N+1}\sum_{i=1}^{N}(x_i - \hat{c}_m^\top z_i)^2. \tag{20}$$

To address permutation invariance in the entries $1, \ldots, m$, we enforce an ordering on the estimated values of C_j, with $C_1 \leq C_2 \leq \ldots \leq C_m$. This breaks the symmetry across the indices of the m new individuals added in the second study, and it is handled by a projection operation at each iteration, in which the estimated values are sorted in ascending order after each maximization step.

Empirical results quantifying the performance of this algorithm are shown in Figure 6, in an experimental setup similar to that for evaluating EM when a single dog is added to a dataset in the main article, with unknown K. A private dataset is assumed to contain 1000 individuals, whereas a separate public dataset of 800 is available; $m = 3$ new individuals are added to the private dataset to produce two parameter vectors β_M and β_{M+m}. On average, the SEM algorithm predicts the correct SNP 75.5% of the time, relative to 71.5% for the ''most common variant'' baseline, a moderate improvement.

5.5. *Scaling of EM algorithm with size of private dataset*

Figure 7 demonstrates the change in accuracy of the EM algorithm over a range of different private database sizes. For this test, a synthetic dataset with 100 SNPs and 1,000,000 individuals is generated; 10,000 are held out as a public database, and 30 individuals are taken as a fixed test dataset of new dogs to add and are used to estimate EM algorithm accuracy, across increasingly large private database sizes. The algorithm has stable performance for increasingly large private databases.

5.6. *Estimating ϕ_0 with different SNP sets*

Here, we analyze what can still be said in the event that the two studies do not use exactly the same set of SNPs. We will still assume that the sets of SNPs are considered to have a significant overlap.

For this purpose we will need a greater variety of notation. A primed variable denotes that it corresponds to the second set of SNPs, for example, K' is the co-occurrence matrix from the original M users for the second experiment. If a vector or matrix is surrounded by square brackets, this denotes the same object but with rows and columns corresponding to SNPs not in the overlap removed, for example, $[K]$ denotes the co-occurrence matrix from the first experiment restricted to the overlapping SNPs.

As described earlier, from the first experiment, we have

$$K\hat{\beta} = \Phi^T y \tag{21}$$

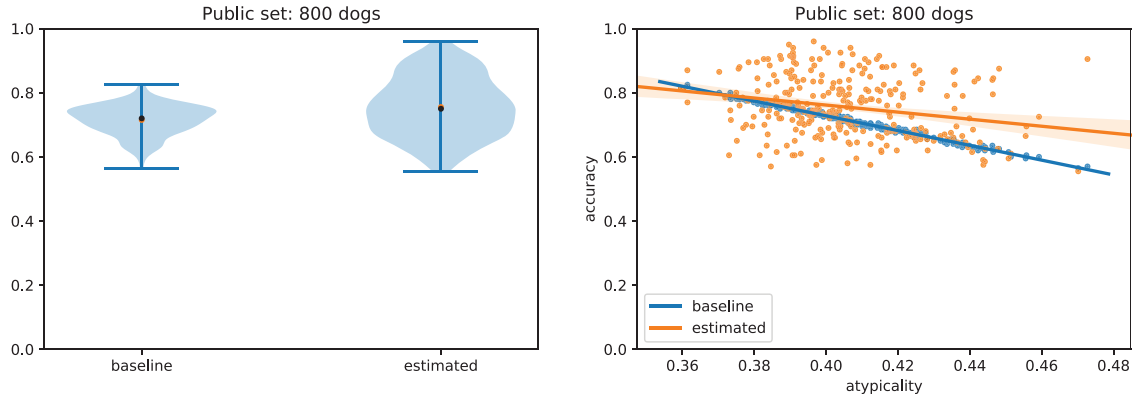


FIG. 6. Results for running the stochastic EM algorithm when estimating SNPs for three additional dogs simultaneously. This experimental setup replicates the experiment for one additional dog, across 5 public/private/test dataset splits, with 20 different test sets of three additional dogs for each. (Left) Accuracy at predicting SNP presence relative to the ''most common variant'' baseline. On average, the SEM algorithm predicts the correct SNP 75.5% of the time, relative to 71.5% for the baseline. (Right) As in the one-dog example, we see relative improvement in the performance of our algorithm when considering more atypical dogs. SEM, stochastic EM.

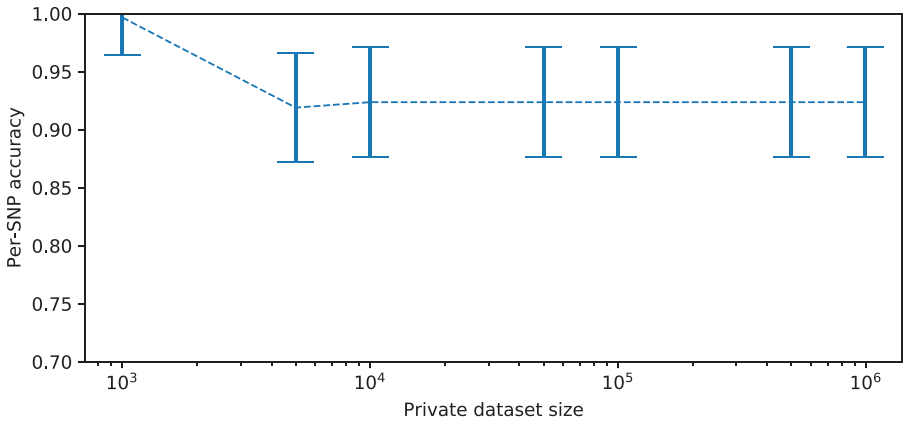


FIG. 7. Accuracy at reconstruction of the genome of one additional individual, using EM estimation and a noisy estimate $\hat{K}$, measured as the size of the initial private database increases. For very small private databases, accuracy is very high, as changes in entries of β are clearly attributable to the new individual. Beyond a certain threshold, overall accuracy is quite stable. Error bars show mean and two standard deviations.

and now, from the second experiment, we have

$$(K' + \phi'^T_0 \phi'_0)\hat{\beta}' = \Phi'^T y'. \tag{22}$$

Taking the difference between these expressions, as earlier, gives

$$K'\hat{\beta}' - K\hat{\beta} = \Phi'^T y' - \Phi^T y - \phi'^T_0 \phi'_0 \hat{\beta}'. \tag{23}$$

Restricting to the overlapping set gives that

$$[K'\hat{\beta}'] - [K\hat{\beta}] = [\Phi'^T y' - \Phi^T y - \phi'^T_0 \phi'_0 \hat{\beta}']. \tag{24}$$

Noting that $[K] = [K']$ and that $[\Phi'^T y'] - [\Phi^T y] = [\phi_0^T y_0]$ we get that

$$[K]([\hat{\beta}'] - [\hat{\beta}]) = [\phi_0^T](y'_0 - \phi_0 \hat{\beta}'). \tag{25}$$

Analogous to the previous cases, $(y'_0 - \phi_0 \hat{\beta}')$ is a scalar that we can label C and we get

$$[\phi_0^T] = \frac{1}{C}[K]([\hat{\beta}'] - [\hat{\beta}]). \tag{26}$$

Thus, if K is known, it can be used to deduce whether the additional individual has each of the SNPs in the overlapping set. If K is not known exactly, it can be estimated from public data just as in the same SNP case.

5.7. Case in which each GWAS study adds two new sets of participants

This article mostly explores the case in which one study's participants are a subset of the other study's participants. Here, we demonstrate that this is equivalent to the case where each of the two studies contains a small number of participants that are not found in the other study.

In particular, let us say that the first study has $M + a$ participants and the second study has $M + b$ participants, where the first M participants are shared between the studies, but there are a participants that are found in the first study but not the second, and b participants that are found in the second study but not in the first. Following on from Eq. (9), we see that:

$$\begin{aligned} K_M(\hat{\beta}_{M+a} - \hat{\beta}_{M+b}) &= K_M(\hat{\beta}_{M+a} - \hat{\beta}_M) - K_M(\hat{\beta}_{M+b} - \hat{\beta}_M) \\ &= \frac{1}{M}\left[\Phi_a^T(y_a - \Phi_a \hat{\beta}_{M+a}) - \Phi_b^T(y_b - \Phi_b \hat{\beta}_{M+b})\right]. \end{aligned}$$

Let us define the following $(N+1) \times (a+b)$ matrix obtained by concatenating the two genotype matrices:

$$\Phi_{a+b} = [\Phi_a, \Phi_b] \tag{27}$$

and the following $a+b$ length vector:

$$r_{a+b} = \left[(y_a - \Phi_a \hat{\beta}_{M+a}),\ -(y_b - \Phi_b \hat{\beta}_{M+b})\right] \tag{28}$$

Then, this gives us:

$$K_M(\hat{\beta}_{M+a} - \hat{\beta}_{M+b}) = \frac{1}{M}\Phi_{a+b} r_{a+b} \tag{29}$$

This means that having two nonoverlapping participant sets is equivalent to the setting in which the first study is a subset of the second (only m is now $a+b$).

5.8. *Description of* K *when the genotypes are non-binary*

In many cases, GRS are calculated on genotype matrices that are non-binary. In particular, they may take on three discrete values 0, 1, and 2, where 0 indicates that the most common variant is homozygous, 1 indicates that the individual is heterozygous for the uncommon variant, and 2 indicates that the individual is homozygous for the uncommon variant.

If this is the case, the description of K will change. However, it is still the case that the entries of K depend only on the SNP frequencies and SNP co-occurrence frequencies in the dataset, and that knowledge of SNP frequencies and pairwise co-frequencies from the original study, are all that is required to compute K.

– For $i = 1, \ldots, N$: $K_{ii} = p_{Aa} + 4p_{AA}$ where p_{aa} is the frequency of individuals being heterozygous for the uncommon variant and p_{AA} is the frequency of individuals being homozygous for the uncommon variant.
– For $i = 1, \ldots, N-1$ and $j > i$: $K_{ij} = K_{ji} = p_{Aa/Bb} + 2p_{AA/Bb} + 4p_{AA/BB}$ where $p_{Aa/Bb}$ is the frequency that both SNPs are simultaneously heterogygous, $p_{AA/Bb}$ is the frequency that one SNP is homozygous for the rare variant and the other is heterogygous simultaneously, and $p_{AA/BB}$ is the frequency that that uncommon variants are found to be homozygous simultaneously.
– For $i = 1, \ldots, N$ and $j = N+1$: $K_{ij} = K_{ji} = p_{Aa} + 2p_{AA}$.
– Finally, $K_{N+1,N+1} = 1$.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

This project was funded by the Alan Turing Institute Research Fellowship under EPSRC Research grant (TU/A/000017); EPSRC/BBSRC Innovation Fellowship (EP/S001360/1), and under the EPSRC grant EP/N510129/1. It was also partly funded by a grant from CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015–2020.

REFERENCES

Belsky, D.W., Moffitt, T.E., Sugden, K., et al. 2013. Development and evaluation of a genetic risk score for obesity. *Biodemography Soc. Biol.* 59, 85–100.

Cai, R., Hao, Z., Winslett, M., et al. 2015. Deterministic identification of specific individuals from GWAS results. *Bioinformatics* 31, 1701–1707.

Celeux, G., and Diebolt, J. 1985. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Stat. Q.* 2, 73–82.

Chouraki, V., Reitz, C., Maury, F., et al. 2016. Evaluation of a genetic risk score to improve risk prediction for Alzheimer's disease. *J. Alzheimers Dis.* 53, 921–932.

Day, F.R., Thompson, D.J., Helgason, H., et al. 2017. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet*. 49, 834–841.

Dwork, C., Smith, A., Steinke, T., et al. 2015. Robust traceability from trace amounts. 2015 IEEE 56th Annual Symposium on Foundations of Computer Science. Berkeley, California. pp. 650–669.

Hayward, J.J., Castelhano, M.G., Oliveira, K.C., et al. 2016. Complex disease and phenotype mapping in the domestic dog. *Nat. Commun.* 7, 10460.

Huang, Z., Ayday, E., Fellay, J., et al. 2015. Genoguard: Protecting genomic data against brute-force attacks. 2015 IEEE Symposium on Security and Privacy. San Jose, California. pp. 447–462.

Im, H., Gamazon, E., Nicolae, D., et al. 2012. On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet*. 90, 591–598.

Jiang, L., Zheng, Z., Qi, T., et al. 2019. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* 51, 1749–1755.

Knowles, J.W., and Ashley, E.A. 2018. Cardiovascular disease: The rise of the genetic risk score. *PLoS Med.* 15, e1002546.

Mittos, A., Malin, B., and Cristofaro, E.D. 2019. Systematizing genome privacy research: A privacy-enhancing technologies perspective. *Proc. Privacy Enhanc. Technol.* 2019, 87–107.

Qi, L., Ma, J., Qi, Q., et al. 2011. Genetic risk score and risk of myocardial infarction in hispanics. *Circulation* 123, 374–380.

Raisaro, J.L., Tramèr, F., Ji, Z., et al. 2017. Addressing beacon re-identification attacks: Quantification and mitigation of privacy risks. *J. Am. Med. Inform. Assoc.* 24, 799–805.

Shringarpure, S., and Bustamante, C. 2015. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* 97, 631–646.

Simmons, S., and Berger, B. 2015. One size doesn't fit all: Measuring individual privacy in aggregate genomic data. 2015 IEEE Security and Privacy Workshops. San Jose, California. pp. 41–49.

Tropp, J.A. 2015. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* 8, 1–230.

Visscher, P.M., Wray, N.R., Zhang, Q., et al. 2017. 10 years of gwas discovery: Biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22.

von Thenen, N., Ayday, E., and Cicek, A.E. 2018. Re-identification of individuals in genomic data-sharing beacons via allele inference. *Bioinformatics* 35, 365–371.

Wang, Y.X. 2018. Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI). Monterey, California. pp. 93–103.

Zhao, X., Xi, B., Shen, Y., et al. 2014. An obesity genetic risk score is associated with metabolic syndrome in Chinese children. *Gene* 535, 299–302.

Address correspondence to:
Dr. Daphne Ezer
Department of Biology
University of York
York YO10 5DD
United Kingdom

E-mail: daphne.ezer@york.ac.uk