White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/161818/

Version: Accepted Version

Proceedings Paper:
Paige, Brooks, Bell, James, Bellet, Aurelien et al. (2020) Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores. In: Schwartz, Russell, (ed.) Lecture Notes in Computer Science:Research in Computational Molecular Biology: 24th Annual International Conference, RECOMB 2020, Padua, Italy, May 10–13, 2020, Proceedings. Lecture Notes in Bioinformatics (LNBI). Springer Nature Switzerland, pp. 266-268.

https://doi.org/10.1007/978-3-030-45257-5_32

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Reconstructing Genotypes in Private Genomic Databases from Genetic Risk Scores*

Brooks Paige[1,2][0000−0002−4797−1563], James Bell[1][0000−0003−4493−4297], Aurélien Bellet[3][0000−0003−3440−1251], Adrià Gascón[1,4], and Daphne Ezer[1,4,5][0000−0002−1685−6909]

[1] The Alan Turing Institute, London, UK
[2] Department of Computer Science, UCL, London, UK
[3] Inria, France
[4] University of Warwick, Coventry, UK
[5] Department of Biology, University of York, York, UK {daphne.ezer}@york.ac.uk

Abstract. Some organisations like 23andMe and the UK Biobank have large genomic databases that they re-use for multiple different genome-wide association studies (GWAS). Even research studies that compile smaller genomic databases often utilise 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 using a largely overlapping set of participants, then 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 analysing 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 SNPs 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: Genomic privacy · Genetic risk scores · GWAS

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

* 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 2015-20

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

to overcome [11]. If an individual's password or ID number gets leaked, then 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. While much of the research focus on long-term privacy of genomic databases rests on the longevity of the encryption scheme [7], it is also important to remember that these genomic databases are not just sitting on a server somewhere, but are being continually utilised 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) [2, 5, 15]. Membership inference attacks have also been developed for the case where a person is allowed to repeatedly query a database to learn if at least one individual contains a particular SNP [13, 14, 16]. These kinds of aggregate statistics about the frequency or presence/absence of a particular SNP 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 — i.e. 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 [18]. Yet, even this information can potentially leak private information about individuals in the database. For instance, [8] 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 manuscript, 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 [1, 3, 4, 10, 12, 20]. 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

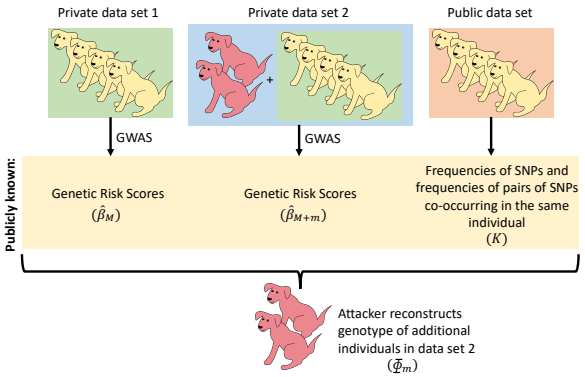


Fig. 1. We investigate the case where two GWAS studies are performed on two data sets that mostly contain the same individuals. We reconstruct the genotype of those individuals added to the second study, using the GRS from each study and an estimate of SNP frequencies.

allow several GWAS to be performed within a short period of time. Under these circumstances, we demonstrate that it is possible to completely reconstruct the SNPs of an individual using a custom Expectation-Maximisation (EM) algorithm. We also provide suggestions for avoiding this kind of attack.

To be clear, this manuscript 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 on 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 (i) how the SNPs were initially filtered or (ii) 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 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).

Afterwards, 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

We also briefly discuss how loosening additional restrictions would impact our ability to predict individual genotypes. In particular, we analyse 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 data sets with largely overlapping populations, it may be possible to reconstruct the genotypes of those individuals who participated in one study but not the other (Figure 1).

2 Methods

Genetic risk score (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, in order to decrease the correlation between them.

In this setting, we suppose 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 a design matrix Φ_M to be the $M \times (N+1)$ matrix

$$\Phi_M = \left[X_M \; 1_M \right]. \tag{1}$$

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

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

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 $\phi_0^\top = [x_0^\top \; 1]$ 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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

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

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

where C is a scalar, specifically $C = \frac{1}{M}(y_0 - \phi_0^\top \hat{\beta}_{M+1})$. This 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 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 we will see 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 a SNP (and the last column contains only 1). The corresponding analog to Eq. (7) for multiple individuals, which we derive in Appendix B, 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 Appendix C 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

- For $i = 1, \ldots, N-1$ and $j > i$: $K_{ij} = K_{ji}$ estimates the probability that SNP i and SNP j are both 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 probability that SNP i has value 1, i.e. $K_{i,N+1} = K_{N+1,i} = K_{ii}$.
- 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 in order 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, while $\hat{\beta}_M$, $\hat{\beta}_{M+1}$ 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, e.g. 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 Appendix D.3, and a formal analysis of the reconstruction error of ϕ_0 given the error in $\hat{K}$ is found in Appendix D.1.

3 Results

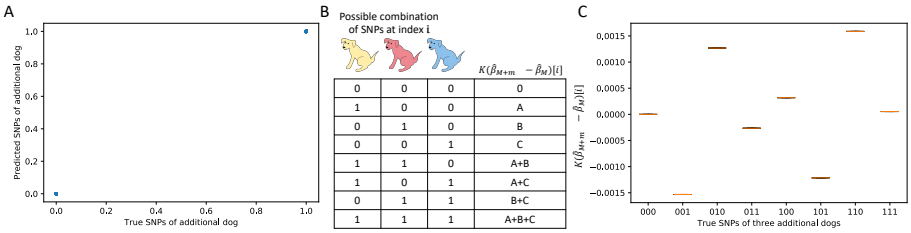
The key observation from the previous section is that the vectors d_1 and d_m, derived from the change in parameter vectors $\hat{\beta}$ from a first study to a second study, *take only a finite number of values* thanks to the fact that the design matrices Φ contain only zeros and ones. In particular, 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 which use the Cornell Dog Genome dataset as a case study [6]. More details on the experimental setup can be found in Appendix A.

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, i.e. 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 Eq. (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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.



0	0	0	0
1	0	0	A
0	1	0	B
0	0	1	C
1	1	0	A+B
1	0	1	A+C
0	1	1	B+C
1	1	1	A+B+C

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 using the data from the Cornell Dog Database, as in (A).

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, e.g. 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, both on common and uncommon SNPs (see Figure 2(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) above. 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 (Figure 2(B)). 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 (see Figure 2(C)).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

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 see the more complete algorithm in Section G.

1. First, extract all unique, non-zero values from d_m.
2. Find the sum of all pairs of values in this vector.
3. Find all values that are in (1), but not in (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 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, both with common and uncommon SNPs (Figure 2(C)).

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. While this could be released voluntarily by organisations which 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.

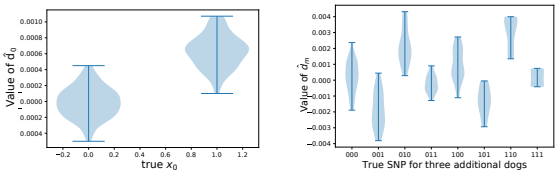


Fig. 3. Example values taken by the noisy vector $\hat{d}$, given the true value of the corresponding SNP in the genome. (Left) adding one new participant; (right) adding three new participants. These figures are analogous to those in Figure 2, albeit in the case where K is not known and instead estimated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

We simulated this scenario 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

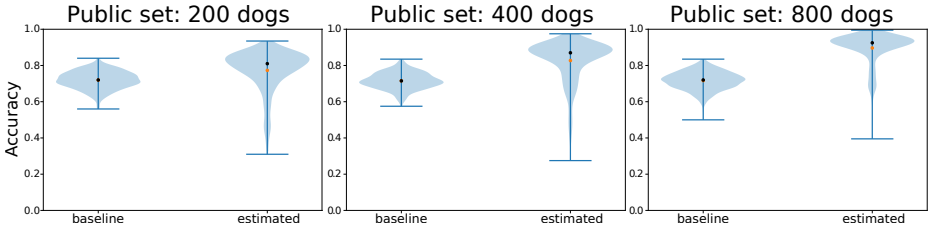


Fig. 4. Accuracy at reconstruction of genomes x_0 using EM estimation and a noisy estimate $\hat{K}$, as compared to a natural baseline which always predicts the most common variant at each SNP locus. We use this as a baseline, because without any additional information about β_M and β_{M+1}, the most 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 (i) took 10 random splits of the full dog data set, assigning dogs to either the public and private data set (ii) 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.

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 expectation-maximisation algorithm to find a maximum likelihood estimate of the constant C and recover ϕ_0, i.e. 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 D.3 for details). We find that this method can successfully reconstruct the correct value of $\phi_0[i]$ much better than a baseline which uses the public dataset to independently estimate the most common variant for each SNPs (see Figure 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 (see Figure 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 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 A. We also derive theoretical error bounds for our estimate of ϕ_0 based on the error in $\hat{K}$ in Section D.1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

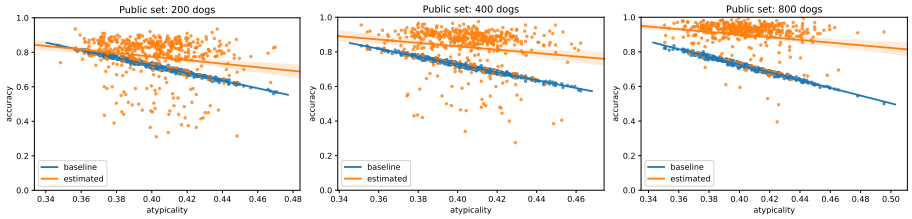


Fig. 5. Results of Figure 4 broken down by individual dogs. Here each point represents a dog and we define *atypicality* as the proportion of *uncommon variants* that the dog has compared to the public database– for instance, if 51% or more of dogs in the public database have a G in a specific locus, but this dog has a T, then this would count towards the dog's atypicality. In other words, dogs further to the right are less and less similar to average dog 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 which are highly dissimilar to those in the public dataset. Larger public databases (right) provide more accurate population estimates $\hat{K}$, leading to more accurate reconstructions overall.

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 (i) significantly correlated to the trait after covariates are considered and (ii) 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 Appendix F.

4 Discussion

In this manuscript, 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 is available to the attacker. For instance, the attacker may have 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

attack and the membership 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, while the attacker may use the reconstructed genomic data for a different (potentially unethical) purpose.

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. However, providing a very noisy $\hat{K}$ may be of limited utility to the scientific community.
2. If a genetic data set is intended to serve for multiple complementary analyses, it is important that all study participants are used in every analysis performed. If there is 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 using the approach in Section 3.2 if the number of SNPs $N > 2^m$. Thus, we suggest to avoid releasing multiple studies which differ by fewer than $\log_2 N$ individuals.

Extensions and future work. While 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 something slightly more complicated that we describe in Appendix H. This does not dramatically change the approach in this paper, except in that the vector d_m can take 3^m possible values, rather than 2^m. In practice, then, studies which 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 in randomly perturbing the GRS models before releasing them, as done in differentially private linear regression [19]. However, a naive application of this strategy could

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

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 paper and left for future work.

Another countermeasure could consist in refraining 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 to. An investigation of how differences in population structures impact K_{est} will be undertaken in public human genomic data sets in future work, but is also beyond the scope of this paper.

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 to test how properties of population structure will impact our ability to estimate K and the accuracy of our reconstruction attack. In addition, the expectation-maximisation algorithm described here is only explicitly described for the case in which one participant is added to the database, and the matrix K is estimated from a public database. We are currently extending the algorithm to the broader case in which multiple participants are added at once.

It may seem on the surface unlikely 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 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 ([9]). In such settings, it could be very possible for multiple released GRS models to be computed on sets of individuals which differ by only a few participants. In future work, we aim to extend our analysis and attack to settings where multiple GRS models are released, each predicting different but highly correlated traits.

References

1. Belsky, D.W., Moffitt, T.E., Sugden, K., Williams, B., Houts, R., McCarthy, J., Caspi, A.: Development and evaluation of a genetic risk score for obesity. Biodemography and Social Biology (2013). https://doi.org/10.1080/19485565.2013.774628
2. Cai, R., Hao, Z., Winslett, M., Xiao, X., Yang, Y., Zhang, Z., Zhou, S.: Deterministic identification of specific individuals from GWAS results. In: Bioinformatics (2015). https://doi.org/10.1093/bioinformatics/btv018
3. Chouraki, V., Reitz, C., Maury, F., Bis, J.C., Bellenguez, C., Yu, L., Jakobsdottir, J., Mukherjee, S., Adams, H.H., Choi, S.H., Larson, E.B., Fitzpatrick, A., Uitterlinden, A.G., De Jager, P.L., Hofman, A., Gudnason, V., Vardarajan, B., Ibrahim-Verbaas, C., Van Der Lee, S.J., Lopez, O., Dartigues, J.F., Berr, C., Amouyel, P.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

Bennett, D.A., Van Duijn, C., Destefano, A.L., Launer, L.J., Ikram, M.A., Crane, P.K., Lambert, J.C., Mayeux, R., Seshadri, S.: Evaluation of a Genetic Risk Score to Improve Risk Prediction for Alzheimer's Disease. Journal of Alzheimer's Disease (2016). https://doi.org/10.3233/JAD-150749

4. Day, F.R., Thompson, D.J., Helgason, H., Chasman, D.I., Finucane, H., Sulem, P., Ruth, K.S., Whalen, S., Sarkar, A.K., Albrecht, E., Altmaier, E., Amini, M., Barbieri, C.M., Boutin, T., Campbell, A., Demerath, E., Giri, A., He, C., Hottenga, J.J., Karlsson, R., Kolcic, I., Loh, P.R., Lunetta, K.L., Mangino, M., Marco, B., McMahon, G., Medland, S.E., Nolte, I.M., Noordam, R., Nutile, T., Paternoster, L., Perjakova, N., Porcu, E., Rose, L.M., Schraut, K.E., Segrè, A.V., Smith, A.V., Stolk, L., Teumer, A., Andrulis, I.L., Bandinelli, S., Beckmann, M.W., Benitez, J., Bergmann, S., Bochud, M., Boerwinkle, E., Bojesen, S.E., Bolla, M.K., Brand, J.S., Brauch, H., Brenner, H., Broer, L., Brüning, T., Buring, J.E., Campbell, H., Catamo, E., Chanock, S., Chenevix-Trench, G., Corre, T., Couch, F.J., Cousminer, D.L., Cox, A., Crisponi, L., Czene, K., Davey Smith, G., De Geus, E.J., De Mutsert, R., De Vivo, I., Dennis, J., Devilee, P., Dos-Santos-Silva, I., Dunning, A.M., Eriksson, J.G., Fasching, P.A., Fernández-Rhodes, L., Ferrucci, L., Flesch-Janys, D., Franke, L., Gabrielson, M., Gandin, I., Giles, G.G., Grallert, H., Gudbjartsson, D.F., Guénel, P., Hall, P., Hallberg, E., Hamann, U., Harris, T.B., Hartman, C.A., Heiss, G., Hooning, M.J., Hopper, J.L., Hu, F., Hunter, D.J., Ikram, M.A., Im, H.K., Järvelin, M.R., Joshi, P.K., Karasik, D., Kellis, M., Kutalik, Z., Lachance, G., Lambrechts, D., Langenberg, C., Launer, L.J., Laven, J.S., Lenarduzzi, S., Li, J., Lind, P.A., Lindstrom, S., Liu, Y., Luan, J., Mägi, R., Mannermaa, A., Mbarek, H., McCarthy, M.I., Meisinger, C., Meitinger, T., Menni, C., Metspalu, A., Michailidou, K., Milani, L., Milne, R.L., Montgomery, G.W., Mulligan, A.M., Nalls, M.A., Navarro, P., Nevanlinna, H., Nyholt, D.R., Oldehinkel, A.J., O'Mara, T.A., Padmanabhan, S., Palotie, A., Pedersen, N., Peters, A., Peto, J., Pharoah, P.D., Pouta, A., Radice, P., Rahman, I., Ring, S.M., Robino, A., Rosendaal, F.R., Rudan, I., Rueedi, R., Ruggiero, D., Sala, C.F., Schmidt, M.K., Scott, R.A., Shah, M., Sorice, R., Southey, M.C., Sovio, U., Stampfer, M., Steri, M., Strauch, K., Tanaka, T., Tikkanen, E., Timpson, N.J., Traglia, M., Truong, T., Tyrer, J.P., Uitterlinden, A.G., Edwards, D.R., Vitart, V., Völker, U., Vollenweider, P., Wang, Q., Widen, E., Van Dijk, K.W., Willemsen, G., Winqvist, R., Wolffenbuttel, B.H., Zhao, J.H., Zoledziewska, M., Zygmunt, M., Alizadeh, B.Z., Boomsma, D.I., Ciullo, M., Cucca, F., Esko, T., Franceschini, N., Gieger, C., Gudnason, V., Hayward, C., Kraft, P., Lawlor, D.A., Magnusson, P.K., Martin, N.G., Mook-Kanamori, D.O., Nohr, E.A., Polasek, O., Porteous, D., Price, A.L., Ridker, P.M., Snieder, H., Spector, T.D., Stöckl, D., Toniolo, D., Ulivi, S., Visser, J.A., Völzke, H., Wareham, N.J., Wilson, J.F., Spurdle, A.B., Thorsteindottir, U., Pollard, K.S., Easton, D.F., Tung, J.Y., Chang-Claude, J., Hinds, D., Murray, A., Murabito, J.M., Stefansson, K., Ong, K.K., Perry, J.R.: Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nature Genetics (2017). https://doi.org/10.1038/ng.3841
5. Dwork, C., Smith, A., Steinke, T., Ullman, J., Vadhan, S.: Robust traceability from trace amounts. 2015 IEEE 56th Annual Symposium on Foundations of Computer Science pp. 650–669 (Oct 2015). https://doi.org/10.1109/FOCS.2015.46
6. Hayward, J.J., Castelhano, M.G., Oliveira, K.C., Corey, E., Balkman, C., Baxter, T.L., Casal, M.L., Center, S.A., Fang, M., Garrison, S.J., Kalla, S.E., Korniliev, P., Kotlikoff, M.I., Moise, N.S., Shannon, L.M., Simpson, K.W., Sutter, N.B., Todhunter, R.J., Boyko, A.R.: Complex disease and phenotype mapping in the domes-

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

tic dog. Nature Communications (2016). https://doi.org/10.1038/ncomms10460
7. Huang, Z., Ayday, E., Fellay, J., Hubaux, J., Juels, A.: Genoguard: Protecting genomic data against brute-force attacks. 2015 IEEE Symposium on Security and Privacy pp. 447–462 (May 2015). https://doi.org/10.1109/SP.2015.34
8. Im, H., Gamazon, E., Nicolae, D., Cox, N.: On sharing quantitative trait gwas results in an era of multiple-omics data and the limits of genomic privacy. The American Journal of Human Genetics **90**(4), 591 – 598 (2012). https://doi.org/https://doi.org/10.1016/j.ajhg.2012.02.008, http://www.sciencedirect.com/science/article/pii/S0002929712000936
9. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., Yang, J.: A resource-efficient tool for mixed model association analysis of large-scale data. bioRxiv (2019). https://doi.org/10.1101/598110
10. Knowles, J.W., Ashley, E.A.: Cardiovascular disease: The rise of the genetic risk score. PLoS Medicine (2018). https://doi.org/10.1371/journal.pmed.1002546
11. Mittos, A., Malin, B., Cristofaro, E.D.: Systematizing genome privacy research: A privacy-enhancing technologies perspective. Proceedings on Privacy Enhancing Technologies **2019**(1), 87 – 107 (2019), https://content.sciendo.com/view/journals/popets/2019/1/article-p87.xml
12. Qi, L., Ma, J., Qi, Q., Hartiala, J., Allayee, H., Campos, H.: Genetic risk score and risk of myocardial infarction in hispanics. Circulation (2011). https://doi.org/10.1161/CIRCULATIONAHA.110.976613
13. Raisaro, J.L., Tramèr, F., Ji, Z., Bu, D., Zhao, Y., Carey, W.K., Lloyd, D.D., Sofia, H., Baker, D., Flicek, P., Shringarpure, S.S., Bustamante, C.D., Wang, S., Jiang, X., Ohno-Machado, L., Tang, H., Wang, X., Hubaux, J.P.: Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. Journal of the American Medical Informatics Association **24**(4), 799 – 805 (2017)
14. Shringarpure, S., Bustamante, C.: Privacy risks from genomic data-sharing beacons. The American Journal of Human Genetics **97**(5), 631 – 646 (2015). https://doi.org/https://doi.org/10.1016/j.ajhg.2015.09.010, http://www.sciencedirect.com/science/article/pii/S0002929715003742
15. Simmons, S., Berger, B.: One size doesn't fit all: Measuring individual privacy in aggregate genomic data. 2015 IEEE Security and Privacy Workshops pp. 41–49 (May 2015). https://doi.org/10.1109/SPW.2015.25
16. von Thenen, N., Ayday, E., Cicek, A.E.: Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics **35**(3), 365–371 (07 2018). https://doi.org/10.1093/bioinformatics/bty643, https://doi.org/10.1093/bioinformatics/bty643
17. Tropp, J.A.: An introduction to matrix concentration inequalities. Foundations and Trends® in Machine Learning **8**(1-2), 1 – 230 (2015)
18. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J.: 10 years of gwas discovery: Biology, function, and translation. The American Journal of Human Genetics **101**(1), 5 – 22 (2017). https://doi.org/https://doi.org/10.1016/j.ajhg.2017.06.005, http://www.sciencedirect.com/science/article/pii/S0002929717302409
19. Wang, Y.X.: Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI). pp. 93–103 (2018)
20. Zhao, X., Xi, B., Shen, Y., Wu, L., Hou, D., Cheng, H., Mi, J.: An obesity genetic risk score is associated with metabolic syndrome in Chinese children. Gene (2014). https://doi.org/10.1016/j.gene.2013.11.006

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

A Experimental details

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 binarised 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% to 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.

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 1 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 refer to as $\hat{K}$. Now we could calculate $\hat{K}(\hat{\beta}_{M+1} - \hat{\beta}_M)$ and compare this to 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 which are each individually considered as candidates for the $(M+1)^{\text{th}}$ dog added to the private dataset.

B Adding multiple dogs

Here we explain Equations (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:

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

This equation can be rearranged to give

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

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

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

This manuscript 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 contain 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 the first. Following on from Equation 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:

$$\Phi_{a+b} = [\Phi_a, \Phi_b] \tag{10}$$

and the following $a + b$ length vector:

$$r_{a+b} = \left[(y_a - \Phi_a\hat{\beta}_{M+a}), -(y_b - \Phi_b\hat{\beta}_{M+b})\right] \tag{11}$$

Then this gives us:

$$K_M(\hat{\beta}_{M+a} - \hat{\beta}_{M+b}) = \frac{1}{M}\Phi_{a+b}r_{a+b} \tag{12}$$

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

D Estimating K

If the true matrix K is unknown, it can be estimated with public data. We denote this estimator by $\hat{K}$. In order 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

evaluate whether the estimate is good that assessment should account for this systematic error as well.

In the following we are primarily concerned with the error due to the subsampling in both the private and public data sets. For convenience we only consider the case of adding a single individual, though the generalization is quite straightforward.

D.1 Analytic bound on $\|\phi_0 - \hat{\phi}_0\|$

If $\hat{K}$ is substituted for K in our reconstruction equation (7) we get an approximation of ϕ_0 which we denote $\hat{\phi}_0$. We would like to bound the (relative) error between ϕ_0 and $\hat{\phi}_0$. In the following, 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 to 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

$$\frac{\|\hat{\varphi}_0 - \varphi_0\|}{\|\varphi_0\|} \leq \|K^{-1}\|\|K - \hat{K}\|, \tag{13}$$

This means we can bound the error by two quantities. The term $\|K^{-1}\|$ is bounded above by $1/\min(\mathrm{eig}(K))$, which is finite as soon as K is non-singular. 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 (13) 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 [17] 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.

D.2 Modelling the error in $\hat{K}$

In this section we define a model to capture the error in $\hat{K}$, which leads to the expectation maximization algorithm for estimating ϕ_0 which is used in the experiments.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

As our estimated $\hat{K}$ drifts from the true K, this expression $\hat{K}(\hat{\beta}_{M+1} - \hat{\beta}_M)$ would produce a wider range of values than just 0 and C. Let $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ be independent noise, which we assume corrupts each element of K_{ij}; i.e. given the estimated matrix $\hat{K}$, suppose

$$K_{ij} \sim \mathcal{N}(\hat{K}_{ij}, \sigma^2), \tag{14}$$

for some small σ^2. This is clearly an oversimplification (as we know K is e.g. bounded and symmetric), but 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

$$\Delta = \hat{\beta}_{M+1} - \hat{\beta}_M \tag{15}$$

which corresponds to the difference between the two GRS model parameter vectors. Given the true value of K, the system of equations

$$C\phi_0 = K\Delta$$

relates the known quantity Δ and the Gaussian-distributed K with the unknown value of C and of the vector ϕ_0.

If K is Gaussian (following Eq. (14)), 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 $K\Delta = C\phi_0$, we can write this as

$$\hat{K}\Delta \sim \mathcal{N}(C\phi_0, \sigma^2 \Delta^\top \Delta I) \tag{16}$$

where C 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_0, which are zeros and ones, as Bernoulli distributions, whose prior probabilities correspond to the public dataset estimated frequencies. This suggests a model for $\hat{K}\Delta$ which is akin to a constrained mixture of Gaussians.

D.3 Derivation of EM algorithm for reconstructing an individual's genotype

Given the model in Eq. (16), we can define an EM algorithm for maximum likelihood estimation of C and σ^2, which then permits easy inference for each entry of ϕ_0.

Denote the entries of ϕ_0 as $z_1, \ldots, z_{N+1}$; denote their prior probabilities as $\alpha_1, \ldots, \alpha_{N+1}$, where $\alpha_1, \ldots, \alpha_N$ are the (public) population frequencies for each

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

SNP, and $\alpha_{N+1} = 1.0$. Let $x_1, \ldots, x_M$ denote the entries of the fixed (observed) vector $x = \hat{K}\Delta$, which is distributed

$$p(x|C, z, \sigma^2) = \mathcal{N}(x|Cz, \sigma^2 \Delta^\top \Delta I).$$

Supposing we know values of C, σ^2, to estimate the entries of ϕ_0 we want to find $p(z|x, C, \sigma^2)$,

$$p(z|x, C, \sigma^2) \propto p(x|C, z, \sigma^2)p(z)$$

For each z_i, $i = 1, \ldots, N$, we define the quantity

$$\pi_i = p(z_i = 1|x, C, \sigma^2) = \frac{\alpha_i \mathcal{N}(x_i|C, \sigma^2 \Delta^\top \Delta)}{\alpha_i \mathcal{N}(x_i|C, \sigma^2 \Delta^\top \Delta) + (1 - \alpha_i)\mathcal{N}(x_i|0, \sigma^2 \Delta^\top \Delta)}, \tag{17}$$

the posterior probability of each particular entry taking a value of 1, rather than 0. A maximum likelihood algorithm to estimate C, σ^2 proceeds by alternately:

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

To maximize C and σ^2, we first compute the derivatives of

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

We take derivative w.r.t. C,

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

and set equal to zero to find

$$\hat{C} = \frac{\sum_i \pi_i x_i}{\sum_i \pi_i}. \tag{18}$$

For the noise term σ^2, we have

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

which with some algebraic rearrangement becomes

$$\hat{\sigma}^2 = \frac{1}{N \Delta^\top \Delta} \sum_{i=1}^{N} \pi_i (x_i - C)^2 + (1 - \pi_i) x_i^2. \tag{19}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

These updates taken together can be used to define an EM algorithm which 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. $\pi_i \equiv p(z_i = 1|x, \hat{C}, \hat{\sigma}^2) = \dfrac{\alpha_i \mathcal{N}(x_i|\hat{C}, \hat{\sigma}^2 \Delta^\top \Delta)}{\alpha_i \mathcal{N}(x_i|\hat{C}, \hat{\sigma}^2 \Delta^\top \Delta) + (1-\alpha_i)\mathcal{N}(x_i|0, \hat{\sigma}^2 \Delta^\top \Delta)}$,
2. $\hat{C} \leftarrow \dfrac{\sum_i \pi_i x_i}{\sum_i \pi_i}$,
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 both setting to the prior probabilities per-SNP estimated from the public data, as well as to the vector of all zeros (corresponding to a "hard" initialization at the value of the baseline estimate), and found no qualitative difference in performance.

E Scaling of EM algorithm with size of private dataset

Figure 6 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.

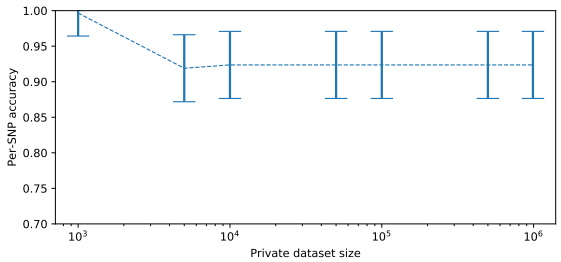


Fig. 6. Accuracy at reconstruction of the genome of one additional individual, using EM estimation and a noisy estimate $\hat{K}$, measured as the size of the initial private database increases. For very small private databases, accuracy is very high, as changes in entries of β are clearly attributable to the new individual. Beyond a certain threshold, overall accuracy is quite stable. Error bars show mean and two standard deviations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

F Estimating ϕ_0 with different SNP sets

Here we analyse 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 the sets of SNPs 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, e.g. 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 that object but with any rows or columns corresponding to SNPs not in the overlap removed, e.g. $[K]$ denotes the co-occurrence matrix from the first experiment restricted to the overlapping SNPs.

As before, from the first experiment, we have

$$K\hat{\beta} = \Phi^T y \tag{20}$$

and now, from the second experiment, we have

$$(K' + \phi_0'^T\phi_0')\hat{\beta}' = \Phi'^T y'. \tag{21}$$

Taking the difference between these expressions, as before, gives

$$K'\hat{\beta}' - K\hat{\beta} = \Phi'^T y' - \Phi^T y - \phi_0'^T\phi_0'\hat{\beta}'. \tag{22}$$

Restricting to the overlapping set gives that

$$[K'\hat{\beta}'] - [K\hat{\beta}] = [\Phi'^T y' - \Phi^T y - \phi_0'^T\phi_0'\hat{\beta}']. \tag{23}$$

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

$$[K]([\hat{\beta}'] - [\hat{\beta}]) = [\phi_0^T](y_0' - \phi_0\hat{\beta}'). \tag{24}$$

Analogously to the previous cases $(y_0' - \phi_0\hat{\beta}')$ is a scalar which we can label C and we get

$$[\phi_0^T] = \frac{1}{C}[K]([\hat{\beta}'] - [\hat{\beta}]). \tag{25}$$

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.

G Algorithm for identifying unique genotypes of multiple dogs, when a precise K is known

While the simple approach described in the main manuscript 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

instance, if there is not a single SNP location where the first individual has a 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, non-zero values from d_m.
2. Find the sum of all pairs of values in (1).
3. Find all values that are in (1), but not in (2).
4. If there are exactly m values in (3) and the sum of these values equal the last value of d_m (corresponding to the intercept term), then you 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 (3) and possibly a few values in (3) that are not in C_m.
6. Begin by subtracting every pair of values in (3). These are now also potential values of C_m
7. Search for a set of m values from (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 (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.

Additionally, 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.

H 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 in order 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.15.907808. this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY 4.0 International license.

- 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, $K_{N+1,N+1} = 1$.