BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

Deterministic Identification of Specific Individuals from GWAS Results

Ruichu Cai [1,2], Zhifeng Hao [1], Marianne Winslett [2,3]*, Xiaokui Xiao [4], Yin Yang [5], Zhenjie Zhang [2]†, and Shuigeng Zhou [6]

[1]School of Computer Science and Technology, Guangdong University of Technology, China.
[2]Advanced Digital Sciences Center, Illinois at Singapore Pte Ltd, Singapore.
[3]Department of Computer Science, University of Illinois at Urbana-Champaign, USA.
[4]School of Computer Science and Engineering, Nanyang Technological University, Singapore.
[5]College of Science, Engineering and Technology, Hamad Bin Khalifa University, Qatar.
[6]School of Computing, Fudan University, China.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The precision of individual genomic information opens the possibility that any participant in genetic databases can be identified despite anonymization strategies. Homer et al.6 showed that information in published genome-wide association studies (GWAS) investigating the relationship between the human genetic variation and disease could lead to the genetic identification of individual participants. Subsequent work, however, showed that though theoretically possible, no practical attack algorithm has been successful.

Results: We have derived the first attack algorithm that can successfully identify specific individuals from limited published associations from the Wellcome Trust Case Control Consortium (WTCCC) dataset. For GWAS results computed over 25 or more randomly-selected loci, our attack algorithm always pinpoints at least one patient from the WTCCC dataset. Moreover, the number of re-identified patients grows rapidly with the number of published genotypes. Finally, we describe methods to disable the attack thus providing a security solution enhancing patient privacy.

Availability: Proofs of the theorems and additional experimental results are available at the support online documents. The attacking code is publicly available at `xxxx`. The data set is available at `http://www.wtccc.org.uk/` on request.

Contact: winslett@illinois.edu, zhenjie@adsc.com.sg

1 INTRODUCTION

GWAS (Hunter *et al.*, 2007; Scott *et al.*, 2007; Sladek *et al.*, 2007; Yeager *et al.*, 2007; Zeggini *et al.*, 2007) are widely used to identify loci in the human genome associated with a specific diseases. The basis of these studies is to associate single-nucleotide polymorphisms (SNPs) or genotypes with the disease phenotype in a case-control design (Hunter *et al.*, 2007). Although a scientific article may present GWAS results at low precision (e.g., correlation between genotypes shown only in a heat map), detailed and accurate results are often available upon request. It is standard to protect the privacy of the participating subjects by keeping patient identities confidential.

Since GWAS results are statistical in nature, until recently most researchers believed that it is safe to share and publish such de-identified results. This belief was challenged by recent bioinformatics research (Homer *et al.*, 2008; Wang *et al.*, 2009), which shows that it is theoretically possible to re-identify individual participants using only aggregate genomic data. Notably, Homer *et al.* (2008) describe the first such method based on statistical hypothesis testing. This method requires aggregate information from many genotypes (e.g., tens of thousands) to obtain high confidence regarding an individual's presence in the aggregate. In contrast, a GWAS usually publishes statistics for a much smaller number of genotypes. Therefore, using the approach suggested by Homer et al., access to the whole genotype association dataset would be necessary to accomplish this identification. Access to such complete datasets is restricted and limited only to qualified biomedical researchers with proper vetting (For example, refer to NIH's policy for sharing GWAS data, http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.). Wang et al.(Wang *et al.*, 2009) propose a more ambitious approach that aims to find all genotypes for every patient in the GWAS. This attack, however, rarely succeeds, since the number of unknowns far exceeds the number of known values. In fact, Wang and colleagues Wang *et al.* (2009) reports only one particular synthetic GWAS involving 174 SNPs and 100 patients on which the attack succeeded. A follow-up study (Zhou *et al.*, 2011) tested this attack with GWAS instances from randomly selected sets of SNPs, and did not find any instance on which the attack succeeded. We conclude that none of the existing methods poses a direct threat to GWAS participants' privacy based on the data that is presented in standard publications.

In data security, the development of effective countermeasures requires the identification of a successful attack algorithm. We devised a strong privacy attack on published GWAS results which successfully identified specific patients by using a strategy of constructing deterministic proofs of study inclusion.

*to whom correspondence should be addressed
†to whom correspondence should be addressed

2 METHODS

2.1 Preliminaries and Problem Definition

A typical GWAS recruits two groups of individuals: *cases* (denoted by D^c) and *controls* (denoted by D^t). Cases are patients of the disease under investigation, and controls are similar people without the disease. Usually, each SNP can have two possible alleles, called the *major allele* (i.e., the more common allele on the SNP) and the *minor allele* (the rarer one). Let A denote the major allele and a be the minor allele, $\{AA, Aa, aa\}$ are the three possible genotypes. Aa is usually not distinguished in some genetic models. Among the three basic models, AA and Aa are taken as the same in the dominant model, aa and Aa are not distinguished in the recessive model, and only the additive model takes Aa as a individual genotypes. Thus, GWAS with two genotypes is the typical case and is the focus of this work. For the simplicity of presentation, the two genotypes are denoted as 0 (major genotype) /1 (minor genotype).

Suppose that the GWAS results involve d loci of the human genome, denoted as g_1, g_2, $\cdots$, g_d. In genetic model with two genotypes, we represent the genomic information of an individual by a d-dimensional binary vector x_i, in which each binary variable x_{ij} represents the genotype of x_i on g_j. Let N^c and N^t be the total number of cases and controls, respectively. For each genotype g_j, we define the following 4 counts of individuals: n_j^c/m_j^c, number of cases having genotype 0/1 on g_j; n_j^t/m_j^t, number of controls having genotype 0/1 on g_j.

A typical GWAS results include N^c, N^t, as well as the following three important statistics: the genotype frequency for each genotype, the genotype-disease association of each genotype, and the genotype-Genotype correlation for each pair of genotypes.

Genotype frequency: The frequency of a minor genotype $g_j = 1$ is usually computed by $F_j = \frac{n^c+n^t}{N^c+N^t}$, i.e., the ratio between the total number of individuals having the minor genotype on g_j, and the total number of GWAS participants.

Genotype-disease association: GWASs commonly use the following equation to measure the association between a genotype (let g_j) and the disease under study: $V_j = \frac{(n_j^c N^t + n_j^c N^c - F_j N^c)^2}{(N^c+N^t-F_j)F_j N^c N^t}$. The asymptotical distribution of V_j is χ^2 distribution with freedom degree 1. The p-value of the insignificance difference is thus $1-\chi^2(d_j, 1)$, in which we abuse $\chi^2(\cdot)$ to denote the accumulative distribution function of χ^2 distribution. In the following, we assume that the GWAS publishes the p-values defined above, denoted as $P(V_j)$ which is true for many GWASs today. Note that our attack is not limited to this particular definition of genotype-disease association, but works on any definition in which n_j^c can be expressed as a function of V_j, N^c, N^t and F_j, such as the one above.

Genotype-Genotype correlation: For each pair of loci (say, g_j and g_k), there are four possible combinations of genotypes, which are (0,0) (i.e., $g_j = 0$ and $g_k = 0$), (0,1), (1,0) and $(1,1)$. Let M_{jk}^{00}, M_{jk}^{01}, M_{jk}^{10} and M_{jk}^{11} denote the number of cases having each of these 4 combinations, respectively. The correlation of g_i and g_j can be measured as follows:$V_{jk} = \frac{(M_{jk}^{11}M_{jk}^{00} - M_{jk}^{10}M_{jk}^{01})^2}{\left(M_{jk}^{11}+M_{jk}^{10}\right)\left(M_{jk}^{01}+M_{jk}^{11}\right)\left(M_{jk}^{00}+M_{jk}^{01}\right)\left(M_{jk}^{00}+M_{jk}^{10}\right)}$. Similar to V_j, V_{jk} also follows the asymptotical χ^2 distribution, with freedom degree 1. Therefore, the corresponding p-value, $P(V_{jk}) = 1 - \chi^2(V_{jk}, 1)$ is published as part of GWAS results.

Algorithm 1 GWAS Attack

Input: D, candidate case set; F_j, frequency of the minor genotypes on g_j; $P(V_j)$, p-value of the association between g_j and the disease; $P(V_{jk})$, p-value of the association between g_i and g_k.

1: Step 1: Recovering the Co-Occurrence Matrix M using F_j, $P(V_j)$ and $P(V_{jk})$ .//
2: Step 2: Finding Presence Proofs ρ using M.//
3: Step 3: Re-Identifying Cases from Candidates D based on proofs ρ.

We assume that the attacker possesses a candidate set D, and the genomic information of each individual in D. By "genomic information" we mean the set of genotypes published in the GWAS results. We distinguish two situations: the first is when the candidate set D is a superset of the GWAS cases D^c, i.e., $D \supseteq D^c$. Such a candidate set can be obtained, for instance, by a curious staff member of the hospital or research center where the GWAS was conducted. Note that the candidate set D can contain much more people that the GWAS cases D^c, e.g., D can be the set of all individuals whose genome sequences are stored in the hospital or research center. This assumption is summarized as follows. The role of this assumption will be discussed in the support online documents.

ASSUMPTION 1. ***Containment***
The adversary knows a candidate set D that contains all the case samples, i.e. $D^c \subseteq D$.

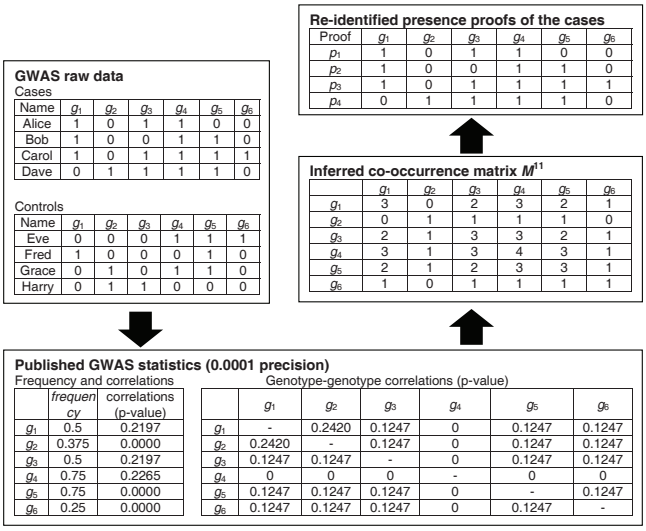
Given the above assumption, the goal of the attacker is to identify individuals in D that belong to the cases of the GWAS based on the GWAS statistics. The formal definition of the attack problem is summarized as follows.

DEFINITION 1. ***GWAS Privacy Attack Problem***
Given candidate set D and GWAS statistics identify as many samples in D as possible that belong to D^c.

2.2 Framework

Assume that the study has identified a set of loci that are associated with the disease, and a set of statistics of the genotypes are published on the genotypes. The published statistics includes, the frequency of the minor genotypes on each identified locus; the p-value of the genotype-disease association for each identified locus; and the p-value of the genotype correlation for each pairs of identified loci. Based on the published statistics, a three step framework is devised to identify specific individuals from GWAS results, recovering the co-occurrence matrix M, finding presence proofs ρ, and re-identifying cases from candidates. The framework is summarized in the Algorithm 1.

Figure 1 presents an example of GWAS results and an overview example of the attack. As shown in the left part of Figure 1, The study has identified a set of loci, and for each locus the GWAS publishes its minor genotype frequency, the correlation between two loci events and their genotype-genotype associations which provides a measure of their correlation. As shown in right part of Figure 1, our privacy attack attempts to reverse the above process. The attack



GWAS raw data

Cases

Name	g_1	g_2	g_3	g_4	g_5	g_6
Alice	1	0	1	1	0	0
Bob	1	0	0	1	1	0
Carol	1	0	1	1	1	1
Dave	0	1	1	1	1	0

Controls

Name	g_1	g_2	g_3	g_4	g_5	g_6
Eve	0	0	0	1	1	1
Fred	1	0	0	0	1	0
Grace	0	1	0	1	1	0
Harry	0	1	1	0	0	0

Re-identified presence proofs of the cases

Proof	g_1	g_2	g_3	g_4	g_5	g_6
p_1	1	0	1	1	0	0
p_2	1	0	0	1	1	0
p_3	1	0	1	1	1	1
p_4	0	1	1	1	1	0

Inferred co-occurrence matrix M^{11}

	g_1	g_2	g_3	g_4	g_5	g_6
g_1	3	0	2	3	2	1
g_2	0	1	1	1	1	0
g_3	2	1	3	3	2	1
g_4	3	1	3	4	3	1
g_5	2	1	2	3	3	1
g_6	1	0	1	1	1	1

Published GWAS statistics (0.0001 precision)

Frequency and correlations

	frequency	correlations (p-value)
g_1	0.5	0.2197
g_2	0.375	0.0000
g_3	0.5	0.2197
g_4	0.75	0.2265
g_5	0.75	0.0000
g_6	0.25	0.0000

Genotype-genotype correlations (p-value)

	g_1	g_2	g_3	g_4	g_5	g_6
g_1	-	0.2420	0.1247	0	0.1247	0.1247
g_2	0.2420	-	0.1247	0	0.1247	0.1247
g_3	0.1247	0.1247	-	0	0.1247	0.1247
g_4	0	0	0	-	0	0
g_5	0.1247	0.1247	0.1247	0	-	0.1247
g_6	0.1247	0.1247	0.1247	0	0.1247	-

Fig. 1. Example of GWAS result publishing and the privacy attack. Top left: part of the raw data of the GWAS, which contains genomic sequences for study participants. Bottom: published results of the GWAS, which lists the genotypes of interest, their frequencies and correlation with the disease, as well as the correction between each pair of these genotypes. Right column: the proposed privacy attack, which first recovers a co-occurrence matrix from the published statistics (only M^{11} is given for space limitation), and uses this matrix to build presence proofs, i.e., sets of genotypes that must be present among the cases.

first infers a co-occurrence matrix M from the published statistics, which contains aggregate information about the cases in the GWAS (only the M^{11} is given in the Figure for the space limitation). Then, the attack applies an iterative data mining algorithm on M to recover sets of genotype sequences that must occur in the cases, which we call presence proofs. Each presence proof contains characteristics of an individual's genome who is one of the cases. Finally, given the genotypes of a particular candidate, the attack checks whether that individual is known to be among the cases, by checking whether their genotype match any presence proof. In the following, we elaborate on these three steps.

2.3 Step 1: Recovering the Co-Occurrence Matrix

The co-occurrence matrix M contains four sub matrix M^{00}, M^{01}, M^{10} and M^{11}, which can be recovered in three steps. Firstly, the diagonal value of M^{11} is inferred based on the published minor genotype frequency F_j and the number of cause N^t. Then, the off-diagonal value of M^{11} is estimated based on the diagonal vale of M, $P(V_j)$ and $P(V_{jk})$. Finally, M^{00}, M^{01}, M^{10} and M^{11} are estimated based on M^{11}.

Each diagonal value M^{11}_{ii} of M represents the number of cases with a minor genotype on g_i, i.e. $g_i = 1$. M^{11}_{ii} is derived directly using the frequency of g_i, the number of cases, the number of controls, and the p-value for the correlation between g_i and the disease. This step is trivial if the GWAS results contains the frequency computed on only the cases, as multiplying each F_j with the total number of cases N^c would yield the corresponding value in M^{11}. Hence, we focus on the case where the minor genotype frequency are computed based on all participants of the GWAS. From the published p-value for the genotype-disease association of each locus (say, g_j), we derive the corresponding value of V_j. Then, using V_j, N^c (i.e., total number of cases), N^t (total number of controls) and F_j, we solve $M_{jj} = n^c_j$ from the definition of V_j.

Each off-diagonal value $M^{11}_{ij} (i \neq j)$ represents the number of cases with both a minor genotypes on g_i and a minor genotypes on g_j, i.e. $g_i = 1$ and $g_j = 1$. Once we have M_{ii}, we can trivially infer M_{ij} in a similar way as for the diagonal values of M^{11}, using the correlation between g_i and g_j, M^{11}_{ii}, M^{11}_{jj}, and the number of cases/controls.

Based on the definition of M, we have $M^{01}_{jk} = M^{11}_{kk} - M^{01}_{jk}$, $M^{10}_{jk} = M^{11}_{jj} - M^{11}_{jk}$ and $M^{00}_{jk} = N^c - M^{11}_{jk} - M^{01}_{jk} - M^{10}_{jk}$. Thus, each elements of M^{00}, M^{01} and M^{10} can be estimated from M^{11}.

When the published statistics are exact, all values of M can be computed by solving simple mathematical equations. When these statistics are only available with limited precision, the computation of M is more complicated. Moreover, it is possible that some values in M cannot be uniquely determined. When this happens, we discard all rows and columns of M that contain at least one undetermined value, and proceed with the remaining sub-matrix of M. In the supporting document, we provide a rigorous analysis of the sufficient conditions for co-occurrence matrix recovery, in terms of the precision of the statistics contained in the GWAS results.

2.4 Step 2: Finding Presence Proofs

The second step of the attack uses the inferred co-occurrence matrix to construct presence proofs. A presence proof (or designated as simply "proof") is a set of genotypes such that at least one patient in the cases has exactly these genotypes. The number of genotypes in a presence proof is called the length of the proof. An example length-3 presence proof is $p = \langle g_1 = 1, g_2 = 0, g_3 = 1 \rangle$. We say that an individual X matches a presence proof if, and only if, X's genome contains all the genotypes of the proof. For example, to match the above presence proof p, an individual's genome must have minor genotype (i.e., genotype 1) on g_1 and g_3, and the major genotype (i.e., 0) on g_2. We call the number of cases matching a proof its frequency. The formal definition of presence proof and matching between a presence proof and an individual are given as follows.

DEFINITION 2. ***Presence Proof and Proof Match***
A presence proof is a quintuple $\rho = (s_\rho, I_\rho, A_\rho, l_\rho, u_\rho)$, where $1 \leq s_\rho \leq d$, called the length of ρ, is the number of genotypes involved in ρ, $I_\rho = \{j_1, j_2, \ldots, j_{s_\rho}\}$ are the indices of the involved loci, $A_\rho = \{a_1, a_2, \ldots, a_{s_\rho}\} \in \{0,1\}^{s_\rho}$ are the genotypes of the proofs on the corresponding loci, l_ρ and u_ρ are the lower bound and upper bound on the number of cases matching ρ. An individual x_i matches *a presence proof ρ,* iff, *for each $j \in I_\rho$, the genotype of x_i on g_j is identical to the corresponding genotype in A_ρ.*

Based on the above definitions, this step aims to identify *presence proofs* by iteratively building longer proofs from shorter ones, using a novel algorithm that resembles Apriori (Agrawal *et al.*, 1994), a commonly used data mining algorithm. In the following, we use the notation D^ρ_c to denote the set of GWAS cases that match a proof ρ. Clearly, $l_\rho \leq |D^\rho_c| \leq u_\rho$. Let $\mathcal{L}_s$ (called *length-s proofs*) denote the set of presence proofs we are going to find that involve exactly s genotypes. $\mathcal{L}_1$ and $\mathcal{L}_2$ can be trivially obtained from the co-occurrence matrix. Specifically, there are two length-1 proofs for each locus g_j: (1, $\{j\}$, $\{0\}$, n^c_j, n^c_j) and (1, $\{j\}$, $\{0\}$, m^c_j,

m_j^c). Regarding $\mathcal{L}_2$, for each pair of loci g_j and g_k, there are 4 proofs: $(2, \{j,k\}, \{0,0\}, M_{jk}^{00}, M_{jk}^{00})$, $(2, \{j,k\}, \{0,1\}, M_{jk}^{01}, M_{jk}^{01})$, $(2, \{j,k\}, \{1,0\}, M_{jk}^{10}, M_{jk}^{10})$, and $(2, \{j,k\}, \{1,1\}, M_{jk}^{11}, M_{jk}^{11})$.

We now describe the iterative procedure that builds a proof of length $s+1$ from two proofs of length s. Given two presence proofs ρ and π of length s, i.e. $s_\rho = s_\pi = s$, we say ρ and π share the same prefix, *iff*. (i) I_ρ and I_π share the same first $s-1$ genotypes, (ii) the last genotype in I_ρ is different than the last one in I_π and (iii) A_ρ and A_π share the same first $s-1$ genotypes. A new presence proof σ of length $s_\sigma = s+1$ is constructed by *merging* ρ and π, denoted as $\sigma = \rho \circ \pi$; specifically, I_σ contains all s indices of I_ρ, plus one more which is the last index in I_π; similarly, A_σ contains all s genotypes of A_ρ, as well as the last genotype in A_π.

It remains to compute of l_σ and u_σ, i.e., the lower bound and upper bound on the number of cases matching σ. We first define the *intersection* ξ of ρ and π (denoted as $\xi = \rho \bullet \pi$) as the prefix that ρ and π share in common, i.e, $s_\xi = s-1$, I_ξ consists of the first $s-1$ indices of I_ρ, and A_ξ consists of the first $s-1$ indices of A_ρ. Since ξ is shorter than ρ and π, it must have been generated before ρ and π in our algorithm, which means that the l_ξ and u_ξ are already known. The following lemma shows how to compute l_σ and u_σ. The proof of the lemma is given in the support online documents.

LEMMA 1. *Given presence proofs ρ and π that share the same prefix, their concatenation $\sigma = \rho \circ \pi$, and their intersection $\xi = \rho \bullet \pi$. Let j_ρ and a_ρ be the last index and genotype in ρ, and j_π and a_π be the last index and genotype in π. We have*

$$|D_c^\sigma| \leq \min\{|D_c^\rho|, |D_c^\pi|, M_{j_\rho j_\pi}^{a_\rho a_\pi}\} \quad (1)$$

$$|D_c^\sigma| \geq |D_c^\rho| + |D_c^\pi| - |D_c^\xi|. \quad (2)$$

Accordingly, we have:

$$u_\sigma = \min\left\{u_\rho, u_\pi, M_{j_\rho j_\pi}^{a_\rho a_\pi}\right\} \quad (3)$$

$$l_\sigma = l_\rho + l_\pi - u_\xi \quad (4)$$

We summarize the presence proof generation procedure in Algorithm 2. The algorithm first generates presence proofs of length 1 and 2 from the co-occurrence matrix. Then, it iteratively generates new proofs of length $m+1$, by concatenating two proofs of length m that share the same prefix.

Figure 2 presents an example of the iterative generation of proofs, for the GWAS data shown in Figure 1. We start with length-1 proofs, which are single genotypes on the loci included in the co-occurrence matrix M. In our example, there are 12 such proofs, i.e., genotype 0 and 1 for each of g_1, , g_6. The frequency of each of these proofs is derived directly from the diagonal values of M, e.g., the frequency of $\langle g_1 = 1\rangle$ is $M_{11} = 3$. We discard proofs with zero frequency, e.g., $\langle g_4 = 0\rangle$, as they do not match any patient in the cases. Next, we construct length-2 presence proofs (e.g., $\langle g_1 = 1, g_2 = 0\rangle$), each by combining two length-1 proofs (e.g., $\langle g_1 = 1\rangle$ and $\langle g_2 = 0\rangle$). We compute the frequency of each length-2 proof as well, from the frequencies of length-1 proofs and off-diagonal values of the co-occurrence matrix. For instance, the frequency of $\langle g_1 = 1, g_2 = 0\rangle$ is computed by subtracting M_{12} (the number of cases with genotype 1 on both g_1 and g_2) from the frequency of $\langle g_1 = 1\rangle$. Again, we discard zero-frequency proofs. A proof of length $s \geq 2$ is built by merging (i.e., taking the set union of all genotypes of) two proofs of length $s-1$ that differ in exactly one genotype. For example, $\langle g_1 = 1, g_2 = 0, g_3 = 1, g_4 = 1\rangle$ can be built by merging $\langle g_1 = 1, g_2 = 0, g_3 = 1\rangle$ and $\langle g_1 = 1, g_2 = 0, g_4 = 1\rangle$. In general, the frequency of a proof with length at least 3 cannot be computed directly from the co-occurrence matrix. Instead, we derive a lower bound and an upper bound for each such proof as shown in Lemma 1. We discard proofs with a frequency lower bound of 0, since they might not match any case. The iterative process continues until no additional proofs can be obtained. In the supporting document, we provide an analysis of the probability of obtaining a proof of a given length, based on the characteristics of the genomic domain.

Algorithm 2 GeneratePresenceProofs

Input: M, the co-occurrence Matrix .

```
Build length-1 and length-2 presence proofs L_1 and L_2.
while L_s ≠ ∅ do
    Initialize an empty L_{s+1}
    for each pair of ρ and π in L_s sharing length s−1 prefix do
        Construct a new presence proof σ = ρ ∘ π
        Find the presence proof ξ = ρ • π
        Set j_ρ and a_ρ be the last index and genotype in ρ
        Set j_π and a_π be the last index and genotype in π
        Set l_σ = l_ρ + l_π − u_ξ
        Set u_σ = min{|D_c^ρ|, |D_c^π|, M_{j_ρ j_π}^{a_ρ a_π}}
        if l_σ > 0 then
            Add σ to L_{s+1}
        end if
    end for
    Increment s by 1
end while
Return all presence proofs
```

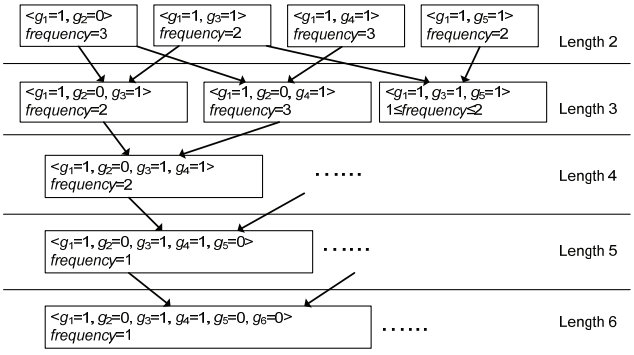


Fig. 2. Presence proofs (length-2 and longer) and their generation. The attack also infers the frequency of each proof. When the frequency cannot be uniquely determined, the attack derives an upper bound and a lower bound for the frequency (as shown for the rightmost proof of length 3). A proof of length l is generated by combining two proofs of length l-1 that differ in exactly one genotype.

2.5 Step 3: Re-Identifying Cases from Candidates

Given the DNA sequence of a suspected study participant, the final step of the attack is to check whether the suspect is among the cases

in the GWAS. From the set of presence proofs obtained in the previous step, we discard each proof that is a subset of another proof. Then we match the suspect's DNA sequence against each proof; if an exact match is found, then we declare the suspect to be a case. When the DNA of the suspect is known before the attack begins, we can significantly speed up processing by only generating proofs that match the suspect's DNA.

Given the set of all presence proofs generated in Step 2, we discard each proof whose genotypes are a subset of another proof's. Among the remaining proofs, we retain only those whose upper bound and lower bound are both 1, i.e., each of them matches exactly one case. The resulting proofs are used to identify cases from candidates. Specifically, if exactly one candidate matches one such proof, we output this candidate as a case in the GWAS. The following theorem ensures that the result of our attack contains no no false positive, under the condition that the target set contains all the cases. The proof of the theorem is provided in the support online documents.

THEOREM 1. *Given that the target set contains all cases, if the proof ρ satisfies $l_\rho = 1$ and there is only one matching individual in the target set, then this individual must be a case in the GWAS.*

Although Algorithm 2 is capable of generating all presence proofs, its computational costs might be too high for a GWAS that publishes a large number of genotypes, due to the exponential number of possible combinations. In the following, we present an optimized algorithm (which we call *candidate matching*) that only generate necessary presence proofs for a given candidate set, rather than enumerating all proofs.

Candidate matching is accomplished by building an appropriate first layer $\mathcal{L}_2$ (originally done on the first line of Algorithm 2), based on the target candidate sample x_i as Formula 5. The rest of the algorithm runs in exactly the same way as Algorithm 2 does.

$$\mathcal{L}_2 = \left\{ \left(1, \{j,k\}, \{x_{ij}, x_{ik}\}, M_{jk}^{x_{ij}x_{ik}}, M_{jk}^{x_{ij}x_{ik}} \right) \right\} \quad (5)$$

To reduce the computation cost, the candidate matching method finds discriminative genotypes before the generation of presence proofs. As the appearance frequency of proofs on the samples is mostly dependent on the co-occurrence counts of the genotypes, we run the genotype selection based on the heuristic that a genotype is more discriminative if the numbers of co-occurrence of the genotype together with other genotypes are consistently smaller. This brings us the simple genotype selection strategy working as follows. Firstly, we calculate the genotype co-occurrence for each genotype in the candidate. If it is a major genotype, we have $h_j = \sum_{1 \le k \le d, k \neq j} \left(M_{jk}^{00} + M_{jk}^{01} \right)$, otherwise, we have $h_j = \sum_{1 \le k \le d, k \neq j} \left(M_{jk}^{10} + M_{jk}^{11} \right)$. The algorithm the returns genotypes with minimal h_j and feed these genotypes to the proof generation procedure.

3 RESULTS AND DISCUSSIONS

To evaluate the effectiveness of the privacy attack, we tested it on eight datasets from the Wellcome Trust Case Control Consortium (WTCCC). All DNA samples in these datasets are collected using the 500K Affymetrix chip, and each sample contains genomic sequence on 394,747 loci. In Table 1, we list the abbreviation, the target disease and the number of cases in each dataset. We simulate seven different GWASs by using the NBS dataset as the controls and one of the seven other datasets as the cases. The reference population is the set of individuals that appear in any of the eight datasets. In each simulated GWAS, we pick a certain number of genotypes uniformly at random, and publish the p-values of their genotype-disease correlations and the correlations between each pair of these genotypes. By default, each published value has a precision of 0.001, and we test different levels of precision in the experiments.

Table 1. WTCCC datasets used in the experiments

Dataset	Case/Control	Disease	Num. of Patients
HT	Case	Hypertension	1952
BD	Case	Bipolar Disorder	1868
CAD	Case	Coronary Artery Disease	1926
CD	Case	Crohn's Disease	1748
RA	Case	Rheumatoid Arthritis	1860
T1D	Case	Type 1 Diabetes	1963
T2D	Case	Type 2 Diabetes	1924
NBS	Control	None	1458

Table 2. Experimental setup for the GWAS simulations. To evaluate the effect of the attack, the experiments vary the number of published loci, the number of loci used in the attack, and the precision of the statistics published by the GWAS.

Parameter	Values of the parameters
Case Dataset	**HT**, BD, CAD, CD, RA, T1D, T2D
Control Dataset	**NBS**
Num. of loci used in the attack	10, 12, **14**, 16, 18
Num. of published loci	25, 50, **75**, 100, 125
Precision	0.1, 0.01, 0.001, **0.0001**, 0.00001

Default values of the parameters in bold

For computational efficiency, we select a subset of the published loci with minimal mutual co-occurrences, and run the privacy attack on this subset. To evaluate the accuracy of the attack, we iteratively consider each member of the reference population as a suspect. We label the suspect as a positive result if at least one presence proof is found in the DNA of the suspect, but nowhere else in the reference population. Otherwise the result is negative, meaning that the suspect was not re-identified as being among the cases. Intuitively, the attack is effective if it returns positive results for the cases and negative answers for other members of the reference population. We repeat the GWAS simulation and attack for 10 different randomly selected sets of published genotypes for each dataset, and report the average results. Table 2 summarizes the parameters investigated in the experiments.

Figure 3 shows that on average, the attack successfully re-identifies 15 cases when 75 genotypes are involved in the GWAS results, of which which just 14 are exploited in the attack. In other words, 14 genotypes out of 75 suffice to find unique patterns in 1% of the cases, patterns that distinguish them from everyone else in the

reference population. Just as importantly, the attack does not falsely re-identify anyone from the reference population. In the supplementary document, we prove that when the reference dataset includes all cases, then the attack will not incur any false positive. We also show that when this assumption does not hold, e.g., some of the cases are not in the reference dataset, false positives are theoretically possible, but unlikely.

Figure 3 also shows that the number of re-identified cases grows rapidly as the number of published genotypes increases. This is important because today's GWAS studies already typically report more than 100 loci in the publication (Sladek *et al.*, 2007; Zeggini *et al.*, 2007), which would tend to boost the re-identification rate significantly. Meanwhile, when the number of published genotypes is fixed, the number of re-identified cases increases with the number of loci used in the attack, at the expense of computation time. In addition, Figure 3 also contains results with varying levels of precision for each published value in the GWAS results. As long as the precision remains above 0.001, the number of re-identified cases tends to be stable; in contrast, when the precision level falls below 0.01, the attack is unable to re-identify any case.

To simulate a real attack, we also tested the effectiveness of our attack on the WTCCC dataset with the genotypes published by Scott *et al.* (2007). Due to the different source of the DNA data employed by Scott et al., only 36 out of the 306 genotypes discussed in their paper are available in the WTCCC datasets. We therefore applied our attack to these 36 genotypes, using the T2D dataset as cases, NBS as controls, and the other six datasets as the reference population. As shown in Figure 4, the attack determines that 12 people from the WTCCC datasets are among the T2D cases, using 14 genotypes that the attack selected from among the 36 available. The attack did not mistakenly re-identify anyone from the reference population as being among the cases. The number of re-identifications is only slightly lower than that achieved with twice as many randomly selected genotypes in Figure 3, further confirming the effectiveness of the attack.

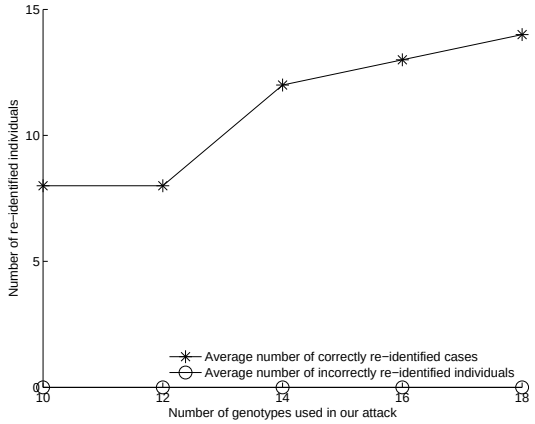


Fig. 4. The number of re-identified cases from the T2D dataset, based on the 36 SNPs published in Ref. (2) that are also available in the WTCCC dataset. The attack re-identifies a dozen cases on average, which is slightly fewer than when the published data is for 75 randomly-selected genotypes. The number of re-identified cases gradually grows when more genotypes are used in the attack. The supplementary document contains additional results obtained by running the same experiment on other datasets of WTCCC.

4 CONCLUSION

To sum up, the privacy attack described in this paper poses a potential threat to the privacy of patients participating in a GWAS. One effective countermeasure is to lower the precision of the published statistics, e.g., publish only a heat map for the correlation between different genotypes, and never reveal their precise values. Meanwhile, since the attack's power grows with the number of genotypes, studies should minimize the number of SNPs included in the published results. Finally, a promising direction for protecting GWAS results with strong privacy guarantees is differential privacy techniques (Johnson and Shmatikov, 2013), which inject random noise into the statistical results. The current state-of-the-art is able to publish a handful of genotypes with the highest correlations with the disease with strong privacy guarantees and good accuracy; however, the method incurs prohibitively high error rates, when a larger number of genotypes are involved in the published results.

With the availability of direct-to-consumer genetic tests that report genotypes associated with medical or physical traits, personal genetic marker data are becoming widely accessible and even public. Medical institutions are considering collecting prospective genomic data on patients in large scale for both research and potentially clinical purposes. It is therefore important for effective security measures to be in place as these data become accessible. We present the first successful attack algorithm using minimum genotype sets and several effective counter-measures. This strategy represents a framework for future genetic privacy defenses.

ACKNOWLEDGEMENT

We thank Anbupalam Thalamuthu for advice on GWAS conventions and trends.

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Funding: R.C. and Z,H. are supported by the National Natural Science Foundation of China (61100148). M.W., X.X., Y.Y. and Z.Z. are supported by SERC 102-158-0074 from A*STAR in Singapore. X.X. is also supported by SUG Grant M58020016 and AcRF Tier 1 Grant RG 35/09 from Nanyang Technological University.

Conflict of Interest: none declared.

REFERENCES

Agrawal, R., Srikant, R., *et al.* (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, **4**(8), e1000167.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., *et al.* (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics*, **39**(7), 870–874.

Johnson, A. and Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1087. ACM.

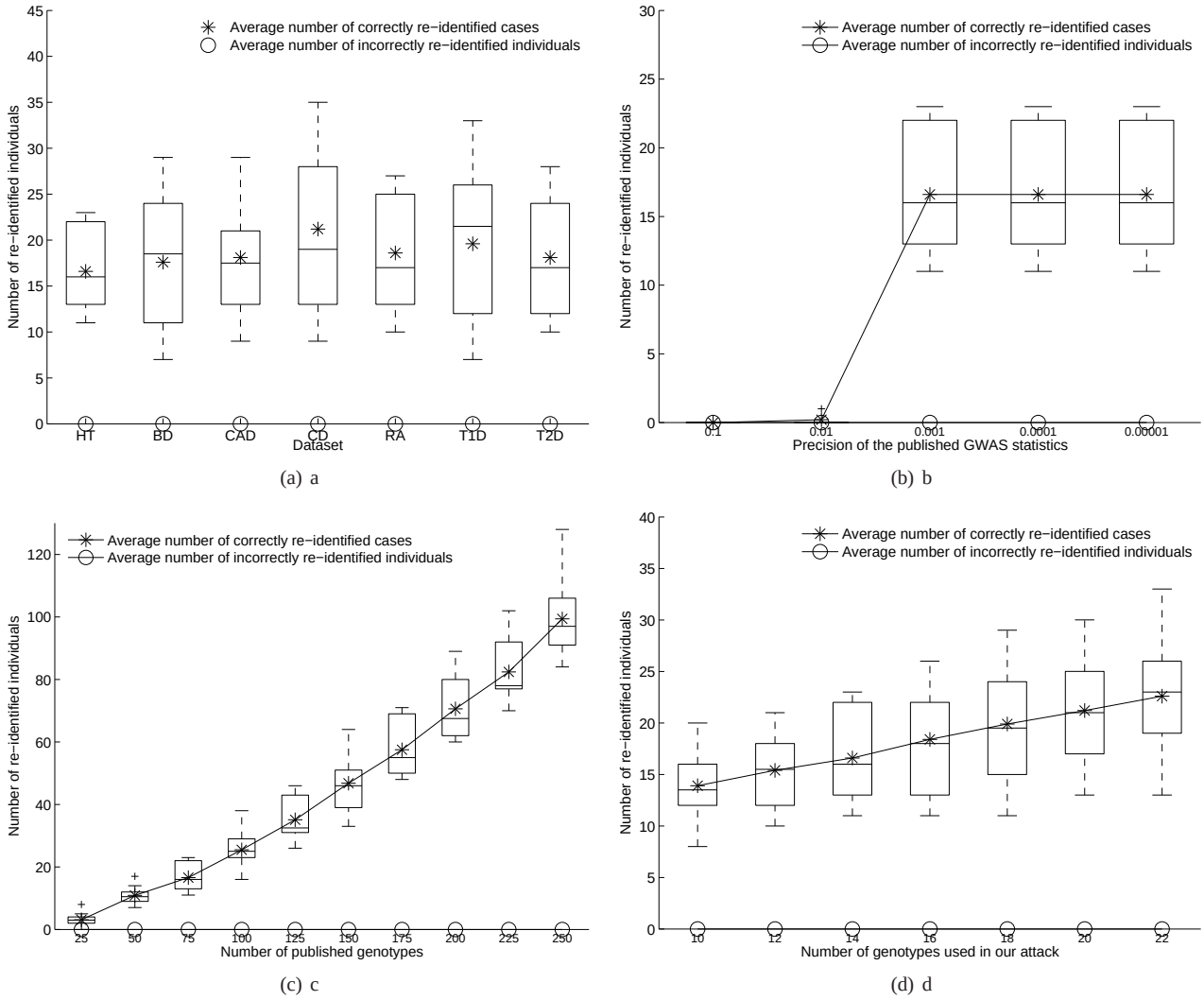


Fig. 3. The number of re-identified cases in the seven WTCCC datasets, averaged across 10 trials with randomly-selected sets of published genotypes. The asterisks show the average number of correct re-identifications. The boxes show the median, 25% quantile, 75% quantile, maximum and minimum numbers of correct re-identifications. Overall, the attack correctly re-identifies at least 10 cases with more than 75% probability, and on average re-identifies 15 cases, which is approximately 1% of all cases. No incorrect re-identifications occurred. (a) Results on the 7 datasets, with default parameter values listed in Table 2. (b) Results with different precisions of the published statistics on the HT dataset, with other parameters fixed to their default values. (c) Results when varying the number of published genotypes on the HT dataset, with other parameters fixed to default values. (d) Results with varying numbers of genotypes used in the attack on the HT dataset, with other parameters fixed to default values. The supplementary document contains additional experimental results.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., *et al.* (2007). A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *science*, **316**(5829), 1341–1345.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., *et al.* (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**(7130), 881–885.

Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. (2009). Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the ACM conference on Computer and communications security*, pages 534–544. ACM.

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., *et al.* (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics*, **39**(5), 645–649.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., *et al.* (2007). Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, **316**(5829), 1336–1341.

Zhou, X., Peng, B., Li, Y. F., Chen, Y., Tang, H., and Wang, X. (2011). To release or not to release: Evaluating information leaks in aggregate human-genome data. In *Proceedings of the ESORICS Conference*, pages 607–627. Springer.