Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study

Rui Wang, Yong Li, XiaoFeng Wang, Haixu Tang, Xiaoyong Zhou
Indiana University Bloomington
Bloomington, IN
{wang63,yonli,xw7,hatang,zhou}@indiana.edu

Abstract

Genome-wide association studies (GWAS) aim at discovering the association between genetic variations, particularly single-nucleotide polymorphism (SNP), and common diseases, which have been well recognized to be one of the most important and active areas in biomedical research. Also renowned is the privacy implication of such studies, which has been brought into the limelight by the recent attack proposed by Homer et al. Homer's attack demonstrates that it is possible to identify a participant of a GWAS from analyzing the allele frequencies of a large number of SNPs. Such a threat, unfortunately, was found in our research to be significantly understated. In this paper, we demonstrate that individuals can actually be identified from even a relatively small set of statistics, as those routinely published in GWAS papers. We present two attacks. The first one extends Homer's attack with a much more powerful test statistic, based on the correlations among different SNPs described by coefficient of determination (r^2). This attack can determine the presence of an individual in a GWAS from the statistics related to a couple of hundred SNPs. The second attack can lead to complete disclosure of hundreds of the participants' SNPs, by analyzing the information derived from the published statistics. We also found that those attacks can succeed even when the precisions of the statistics are low and part of data is missing, which makes the effects of such simple defense limited. We evaluated our attacks on the real human genomes from the International HapMap project, and concluded that such threats are completely realistic.

Categories and Subject Descriptors

K.6.5 [**Security and Protection**]: Unauthorized access

General Terms

Security

Keywords

Genome Wide Association Study, Single Nucleotide Polymorphism, Test Statistics, Markov Model, Integer Programming

CCS'09, November 9–13, 2009, Chicago, Illinois, USA.

1. INTRODUCTION

The rapid advancement in genome technology has revolutionized the field of human genetics by enabling the large-scale applications of genome-wide association study (GWAS) [7], a study that aims at discovering the association between human genes and common diseases. To this end, GWAS investigators determined the genotypes of two groups of participants, people with a disease (cases) and similar people without (controls) in an attempt to use statistical testing to identify genetic markers, typically single-nucleotide polymorphisms (SNP), that are associated to the disease susceptibility genes [50]. If the variation of a SNP is found to be significantly higher in the case group than that in the control group, it is reported as a potential marker of the disease. Of great importance to such a study is the privacy of the participants, whose sensitive information, personally identifiable genetic markers in particular, should not be leaked out without explicit consent. So far, this has been enforced through an informed consent from the participants [10] and an agreement from the investigators to ensure proper use of data according to the consent. Unfortunately, while this process prevents explicit misuse of the participants' genomes, it turns out to be insufficient for deterring information leaks in a more implicit way. Particularly, this paper reports a surprising finding of our research: even the test statistics computed over a small set of SNPs, like those routinely published in GWAS papers, could reveal a substantial amount of genetic information about the participants, and even lead to disclosure of their identities.

The inadequacy of privacy protection in current genome research has also been pointed out by other researchers. For example, Malin et al [49] show that even after removal of explicit identifiers (e.g., name, social security number), an individual could still be identified from a genetic database by examining the genetic markers related to her phenotypes (e.g., eye, skin and hair color). More seriously, Homer et al [42, 65, 5] recently proposes a statistical attack that could determine the presence of an individual in a group (e.g. the cases) from the aggregate *allele frequencies*, i.e. the proportion of all variants occurred at each SNP site, of the whole group. Unlike the threats that have been extensively studied in statistical disclosure control (also inference control) [57, 28], Homer's attack takes advantage of the rich background information related to human genome available in the public domain, as well as the particular statistical properties of genomic data: it compares the victim's SNP profile against the allele frequencies of two populations, a "mixture" such as the case group in a GWAS and a reference population that can be acquired from public data sources such as the International HapMap project [8]; given the profile of a sufficient number of independent SNPs from the victim (at least 10,000), her affiliation in the mixture can be determined with high confidence. The impact of this finding is significant. As an example, the NIH

Table 1: GWAS Terminologies used in this paper.

Terminologies	Description
Polymorphism	The occurrence of two or more genetic forms in the same population of a species (e.g. the human being).
Single Nucleotide Polymorphism(SNP)	The smallest possible polymorphism, which involves two forms different at only a single nucleotide (A, T, C or G) in the human genome.
Allele	Each of the two sets of DNAs in a single human individual's genome inherited from one parent. For each SNP site, two most common nucleotides in the population are referred to as the major and the minor alleles, denoted as 0 and 1, respectively.
Genotype	The combination of two set of alleles. For a SNP site with two common alleles in human population, there are three possible genotypes: two homozygotes, 00 and 11, and one heterozygote 01. In average, the genotypes of two human individuals are approximately 99.9% identical.
Locus(plural loci)	The surrounding regions of a SNP site in the genome .
Haplotype	The types of alleles across multiple neighboring SNP sites in a locus, often also referred to as *SNP sequence*. Each individual has two haplotypes in any locus, each corresponding to one allele. Some haplotypes are more common than others in the population.
Linkage disequilibrium(LD)	Non-random association of alleles among multiple neighboring SNP sites.

announced to swiftly remove all aggregate data of GWAS, including allele frequencies, from public websites [65, 5].

Homer's attack made an important step towards better understanding of the privacy risks involved in publishing genomic data. However, its impacts on GWAS remain uncertain and many questions are left to be answered. First, the paper assumes that the distribution of allele frequencies in the mixture is identical to that in a reference population drawn from the HapMap, which may not be true for a GWAS, as the frequency distribution in its case group is typically different from that of the general population. Second, an attempt to identify an individual using the technique would need a highly-dense genomic profile ($> 10,000$ independent SNPs) from the victim as well as their corresponding allele frequencies from the mixture. Finally, most data released by a GWAS are test statistics such as p-values and r square (r^2) instead of allele frequencies, and the privacy implications of these statistics are still unknown.

In this paper, we show that in the absence of proper protection, even a moderate disclosure of those test statistics, as did most GWAS papers, could pose a privacy risk that cannot be ignored. We present two attacks on these statistics. The first attack could statistically identify an individual in the case group from a small set of statistics, which in some cases are only related to a single *locus*, the surrounding region in the genome where a disease susceptibility SNP is discovered in a GWAS. Like Homer's attack, our technique also needs a reference population, which can be obtained from the reports of the same study conducted over other populations[1], and a SNP profile from the victim. Unlike the prior approach, however, our attack utilizes r^2, a measurement of the correlation between SNPs (referred to as the linkage disequilibrium (LD)), which is much more powerful than allele frequencies of individual SNPs. As a result, the presence of an individual can be determined from the statistics at some locus involving only a couple of hundred SNPs. We also demonstrate that some individuals can be identified even in the absence of a good reference. The second attack can be even more powerful: in the case that a GWAS works on *haplotype* sequences, which is increasingly likely with the maturity of genotype phasing techniques [62, 60, 20, 15, 26, 29], our approach could utilize integer programming to recover from the pair-wise correlations of SNPs (measured by r^2 or D') hundreds of the participants' SNPs. We also describe a technique that reverse engineers the statistics (e.g., p-values, r^2 and D') to calculate pair-wise allele frequencies, a necessary step in both of the proposed attacks. This technique works effectively even when the precision of these statistics is low.

[1]This happens frequently in GWAS, which need to verify the findings of one study over other datasets or reuse the same dataset with new data being added.

We believe that our paper makes the following contributions:

•*Novel identification attacks on GWAS statistics*. We developed novel techniques to recover personally identifiable information from the test statistics published by GWAS papers. These techniques are powerful, capable of restoring hundreds of SNPs and identifying an individual using a much smaller number of SNPs than the prior attack [42]. This suggests that privacy threats in genome research are much more realistic than we thought.

•*Understanding of the limitations of simple countermeasures*. Our research shows that mitigation of such threats is anything but trivial. A simple solution like publishing only coarse-grained statistics might not work well, as the relations among various statistics can still give away a sufficient amount of information for restoring fine-grained data. The strength of such a "correlation" attack demands a well-thought-out response that is built upon analysis of the connections among these statistics. An alternative approach, publishing less data, faces the challenges of selecting a right threshold. Failure to do so might still let our attacks succeed.

•*Implementation and evaluations*. We implemented the proposed attacks and evaluated them against human genome data. Our experimental study demonstrates that these attacks can work in realistic environments, extracting personally identifiable information from the amount of data routinely published in GWAS papers.

The rest of the paper has been organized as follows: Section 2 introduces the background knowledge about GWAS and the statistics involved in the studies; Section 3 elaborates the attacks we propose; Section 4 reports the evaluations of our attacks on real genome data; Section 5 discusses the limitations of our techniques and possible defense; Section 6 surveys the related research and Section 7 concludes the paper.

2. GENOME WIDE ASSOCIATION STUDY

In this section, we briefly survey the procedure of a GWAS, the test statistics it uses and a potential attack on it proposed by prior research. For the convenience of the readers not familiar with genomics, we explain in Table 1 the genomic terminologies necessary for understanding our techniques.

2.1 GWAS: Steps and Test Statistics

A GWAS takes multiple steps to unravel the association between genetic variation and a common disease. Researchers first genotype participants from the case and control groups to extract a set of SNP profiles on selected sites. Usually two alleles can be found at each SNP site, referred to as the major and the minor alleles, denoted by 0 and 1. After proper quality control, the allele frequencies of these SNPs (i.e. the frequencies of 1 or 0) are computed over the case and control respectively. These frequencies are then used as inputs to an association test.

Table 2: GWAS Statistics. Here C_{pq} represents the count of an allele pq ($p, q \in \{0, 1\}$).

2×2 contingency table for Disease - SNP association study of two SNPs

SNP	Disease: Control	Disease: Case
0 (Other allele)	C_{00}	C_{01}
1 (Risk allele)	C_{10}	C_{11}

2×2 contingency table for the linkage disequilibrium analysis of two SNPs

SNP 1 Alleles	SNP 2: 0	SNP 2: 1	Sum
0	C_{00}	C_{01}	C_{0*}
1	C_{10}	C_{11}	C_{1*}
Sum	C_{*0}	C_{*1}	

Formula for related measures of association and LD

Statistics	Formula	Asymptotic distribution
log odds ratios	$\log \frac{C_{00}C_{11}}{C_{01}C_{10}}$	Gaussian
r^2	$\frac{(C_{00}C_{11}-C_{01}C_{10})^2}{C_{0*}C_{1*}C_{*0}C_{*1}}$	Chi-square
D'	$\frac{C_{00}C_{11}-C_{01}C_{10}}{D_{max}}$	-

Association tests are used to detect the SNPs significantly associated to the cases (thus potentially to the disease) under the study. Examples of the tests include Pearson's chi square (χ^2) [55], logistic regression [14], Fisher's exact test [35] or Crochnan-Armitage's test for trends [16]. These tests are performed over the allele frequencies of the case and control groups, and their outcomes are used to calculate a p-value for each SNP. The SNPs with sufficiently small p-values, for example, below 10^{-7}, are selected as putative markers. The associations between the markers and the disease can be quantified using odds ratios (mostly with a confidence level of 95%) or log-odds ratios. For each SNP, three odds ratios are computed, for the risk allele, heterozygote and homozygote respectively. Note that SNP-disease associations (the p-values and the odd-ratios) can be evaluated not only at the single SNP genotype level, but also at the level of *haplotype* (allele combinations involving multiple SNPs), which was shown a stronger statistical power than the genotype-based tests. To achieve this, a class of algorithms called the phasing algorithms [2] will be first applied to infer the most likely haplotypes of the individuals (two haplotypes for each individuals) in the case and control group from their genotypes.

Once putative markers have been detected, the study often moves on to map their associations with other SNPs in the same loci, which is referred to as the *linkage disequilibrium* (LD) [56]. This typically involves measurement of pair-wise allele frequencies (the frequencies of specific two-SNP allele combinations), and calculating statistics such as sensitivity indices (D') or coefficients of correlation (r^2) over the frequencies. These statistics can help identify other SNPs also related to the disease.

One last step of GWAS is to replicate the study on other case and control groups to verify whether the association between the SNP markers and the disease, as identified in prior steps, can also be observed from those populations.

Table 2 lists the formula for calculating aforementioned test statistics. These statistics are routinely published in GWAS papers [30, 59, 58, 62, 43]. Typically, p-values of tens or sometimes hundreds of SNPs are reported. Thousands of r^2 or D' that reflect the LD among these SNPs[2] are often illustrated in figures, and can be acquired from authors without any restrictions. Sometimes, detailed accounts of replication studies are also made public, which discloses multiple populations with identical allele-frequency distributions, and can therefore be used in Homer's attack and our attack elaborated in the follow-up sections.

2.2 Homer's Attack

A statistical attack recently proposed by Homer, et al [42] is believed to threaten the privacy assurance in current GWAS. Homer's technique is primarily designed for analyzing genotypes. It can also be applied to the data after phasing [2] in which two allele sequences (each corresponding to one copy of DNAs inherited from one parent) are derived. This process is actually increasingly used in GWAS [62, 60, 20, 15, 26, 29] . For simplicity of presentation, here we describe this attack on phased genotypes, i.e., haplotypes or SNP sequences.

In Homer's attack, the attacker is assumed to already have a high-density SNP profile of the victim, which can be extracted from a small amount of blood sample. This assumption is realistic, as the cost of genotyping is becoming increasingly affordable [4]. What the attacker wants to determine is the presence of an individual in the case group, an indicative of her contraction of a disease[3]. To this end, the attacker measures the distances between the allele frequency of every SNP j on the profile, $Y_j \in \{0, 1\}$, and the corresponding frequencies in the reference Pop_j, and a mixture M_j respectively. These distances are used to compute the following statistic:

$$D(Y_j) = |Y_j - Pop_j| - |Y_j - M_j| \quad (1)$$

Assuming that the distributions of individual SNPs' allele frequencies in the mixture and the reference population are identical, $D(Y_j)$ will have the same distributions across all independent SNP j. As a result, their sum, according to the central limit theorem, will converge to a normal distribution. The mean of the distribution is zero if the victim is not in the case group, and positive otherwise. Using statistical hypothesis testing, the authors found that 25,000 SNPs of a member in a mixture built from the HapMap resulted in a p-value below 10^{-6}, given the null hypothesis that she does not belong to the mixture.

To apply Homer's attack to GWAS, one needs to have a reference population with an identical distribution of allele frequencies to that of case groups. As discussed before, this could be achieved by taking advantage of the outcomes reported by replication studies or other studies on the same disease. A problem here, however, is the requirement for the information of tens of thousands of SNPs from both the case group and the reference population, which is hard to come by from research papers alone. Moreover, GWAS research typically reports test statistics, instead of allele frequencies. A more credible threat, therefore, is expected to work on the statistics and rely on the information related to a relatively small set of SNPs, on the order of tens or hundreds. Such a threat is discovered in our research to be completely realistic. ?

3. OUR ATTACKS

In this section, we elaborate our attacks on GWAS. We start with a technique that finds allele frequencies from statistics, the foundation for the attacks. Then, we present a statistical attack that works on a small set of pair-wise frequencies, and an integer-programming attack that is capable of recovering hundreds of SNPs at some loci. We also discuss the limitations of simple countermeasures to these attacks.

3.1 From Statistics to Allele Frequencies

A SNP contains a major allele, denoted by 0, and a minor allele, 1. Their individual frequencies and the allele frequencies of SNP pairs (00, 01, 10, and 11) contain a large amount of information. The former is the main ingredient of Homer's attack, and both are needed in our attacks, as elaborated in Section 3.2 and 3.3. A

[2]For n SNPs, the r^2 or D' of $\binom{n}{2}$ pair-wise alleles can be calculated.

[3]This apparently can be determined by comparing the victim's profile with the outcomes of the GWAS on the disease. The problem here, however, is that more often than not, GWAS actually reports inconclusive results.

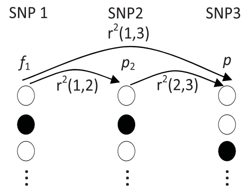

Figure 1: Recover allele frequencies.

GWAS typically reports the allele frequencies for the SNPs identified as putative genetic markers for a disease. For other SNPs, however, only their p-values and LD statistics, r^2 or D', are published. Therefore, the first problem an attacker needs to solve is how to recover the frequencies from publically-available statistics.

This step is by no means trivial. For example, though the mapping between a p-value and the outcome of an association test is typically one-on-one, there could be multiple frequencies that correspond to that outcome. For example, using Pearson's chi square, a p-value of **0.01** was found to associated with 80 possible SNP frequencies in a population studied in our research. The problem is further complicated by the fact that published statistics typically have low precisions, which makes determination of their input frequencies even more difficult.

On the other hand, the statistics published by a GWAS are often connected: for example, the p-values of individual SNPs are bridged by the r^2 between them. Such a connection can be leveraged to reveal their corresponding allele frequencies, which we elaborate below.

Inference of single SNP allele frequencies. We found that the allele frequencies of single SNPs can be restored by "propagating" a genetic marker's frequencies to other SNPs through their linkage disequilibrium, often described by r^2 or D'. Loosely speaking, such a correlation tells us how likely one SNP of an individual can be used to infer some of her other SNPs. The r^2 between two SNPs, as illustrated in Table 2, can also be calculated as follows:

$$r^2 = \frac{(C_{00}N - C_{*0}C_{0*})^2}{C_{0*}C_{1*}C_{*0}C_{*1}} \tag{2}$$

where N is the size of the population, C_{00} is the counts of the pairs of the SNPs' major alleles, C_{0*} and C_{1*} are the counts of the first SNP's major and minor alleles respectively, and C_{*0} and C_{*1} are the allele counts of the second SNP. In the case that the first SNP is actually a marker, whose counts (C_{0*} and C_{1*}) are known, C_{*0} and C_{*1}, which are interdependent, can be determined by the r^2 once C_{00} is known. Actually, C_{00} also relates to C_{*0}, C_{0*} and N. Using these relations as constraints, we can find solutions for C_{*0} and C_{*1} through constraint solving [1]. This can be done efficiently, as the searching space here is bounded by N^2.

It is possible, however, that those constraints are satisfied by more than one solution. To make the solution unique, we need to find more constraints. An immediate one is the p-value of the second SNP, which is computed over the SNP's allele counts in both case and control populations. Other constraints come from the relations among SNPs, as illustrated in Figure 1. Consider three SNPs, S_1, S_2 and S_3, where S_1 is a marker. Besides the aforementioned constraints that exist in pairs (S_1, S_2) and (S_1, S_3), r^2 for (S_2, S_3) can also be leveraged. Similarly, more constraints can be found by looking into the correlations among more SNPs. This approach was demonstrated to be very effective in our research: it completely recovered single SNP frequencies from the statistics with only moderate precisions (r^2 rounded to 2 decimal places, given a population of 200 individuals). Though we discuss our technique here using r^2, the same approach can also be applied to D'.

Recovery of pair-wise frequencies. From the allele frequencies of individual SNPs, pair-wise frequencies can be directly calculated. Specifically, solving Equation 2 with r^2, C_{0*}, C_{1*}, C_{*0} and C_{*1} gives us C_{00}. The counts of other pair-wise alleles, C_{01}, C_{10} and C_{11}, are found from the following linear equations:

$$\begin{aligned} C_{0*} &= C_{00} + C_{01} \\ C_{1*} &= C_{10} + C_{11} \\ C_{*0} &= C_{00} + C_{10} \\ C_{*1} &= C_{01} + C_{11} \end{aligned} \tag{3}$$

Inaccurate statistics. A practical hurdle for our attacks is that the published statistics are often of low-precision. This can be handled by changing Equation 2 to an inequality, giving r^2 a range of acceptable values. Specifically, we use the following constraint to find single-SNP allele counts:

$$L < \frac{(C_{00}N - C_{*0}C_{0*})^2}{C_{0*}C_{1*}C_{*0}C_{*1}} < U \tag{4}$$

where L and U are the lower and upper limits of r^2 respectively.

A more serious problem is that in some cases, published r^2 can be inconsistent with each other, and with population size and p-values. The problem happens in the GWAS that does not phase genotypes and instead, uses maximum likelihood estimation (MLE, e.g., function `LD()` in R [11]) to estimate a pair-wise allele frequency independently from other frequencies. A solution can be using integer programming to estimate a set of frequencies that minimizes the distance between the r^2 computed from them and published statistics. This treatment could bring in some incorrect frequencies that might impact on the follow-up analysis. Fortunately, with the development of phasing techniques nowadays, more and more GWAS works on haplotype [9] and therefore reports the statistics computed from consistent pair-wise frequencies.

Signs. An important piece of information for our statistical attack (Section 3.2) is *sign*, which is determined by the equality $C_{00}C_{11} > C_{01}C_{10}$: it is positive if the inequality holds, and negative otherwise. It is conceivable that signs are much easier to recover than allele frequencies, which can actually be used to compute the signs. In our research, we first ran our constraint solver on $C_{00}C_{11} > C_{01}C_{10}$, together with other constraints for inferring frequencies, and then on $C_{00}C_{11} < C_{01}C_{10}$. The sign is recovered if no solution is found in one of these two cases.

3.2 A Statistical Attack

We follow the strategy proposed by Homer et al. to design our statistical attack. The goal of our attack is to determine the likelihood of a victim to be in a case group of a GWAS study based on one given SNP sequence of a victim. To achieve this goal, we first establish a *reference* group (Figure 2), consisting of the SNP sequences from a group of individuals, drawn from a reference population with the same genetic background as the case group. The International HapMap project [8] provides a large source of samples for this exercise, containing individuals from various ethnic groups, including Nigeria (Yoruba), Japan/China and US residents with ancestry from Northern and Western Europe. In Section 4.2, we will show that the statistical power of our attack relies on the selection of the reference group. However, even when the reference group does not completely mimic the genetic background of the case group, the attack still works, although with a lower power.

Once the reference group is established, we propose a hypothesis test on the SNP sequence of the victim to determine her presence in the case group. A high confidence estimated by this test indicates that the victim's SNP sequence is significantly closer to the reported LD structure (measured by the pair-wise r^2 values) than any other sequence in the reference group, and it is unique enough for identifying the victim. Below we describe this test statistic in details.

Test statistic. Given N sequences of M neighboring SNPs in the genome, we define the *signed* allele correlations r_{ij} between two SNPs i and j ($1 \le i < j \le M$) as $r_{ij} = \frac{C_{11}C_{00} - C_{01}C_{10}}{\sqrt{C_{1*}C_{0*}C_{*1}C_{*0}}}$, where C_{pq} is the pair-wise allele counts, i.e. the number of haplotypes with allele p ($p \in \{0,1\}$) at SNP i and allele q ($q \in \{0,1\}$) at SNP j, and $C_{p*} = C_{p0} + C_{p1}$ and $C_{*q} = C_{0q} + C_{1q}$.

The signed allele correlations can be computed solely from a set of given SNP sequences, and thus can be easily computed from the reference group (denoted as r^R, Figure 2). On the other hand, although the signed allele correlations of the case individuals (denoted as r^C) usually are not reported in GWAS papers (they publish r^2 instead), most (if not all) of them can be derived from the reported data, i.e., r^2 with signs recovered from constraint solving (Section 3.1). To test on the SNP sequence of the victim, $H = h_1h_2...h_m$ (where $h_i \in \{0,1\}$ is an allele), we first transform it into pair-wise indicators, Y_{ij}^{pq}: for any pair of SNPs i and j and pair of alleles p and q ($p, q \in \{0,1\}$), if $h_i = p$ and $h_j = q$, $Y_{ij}^{pq} = 1$; otherwise, $Y_{ij}^{pq} = 0$. Note that for a specific pair of SNPs, there is one $Y^{pq} = 1$ and three $Y^{pq} = 0$ among four possible allele pairs (i.e. 11, 10, 01 and 00). Now we are ready to define the hypothesis test statistic, T_r as,

$$\begin{aligned} T_r &= \textstyle\sum_{1 \le i < j \le N} T_{ij} \\ &= \textstyle\sum_{1 \le i < j \le N} (|(Y_{ij}^{00} + Y_{ij}^{11}) - (r_{ij}^R + 1)/2| \\ &\quad -|(Y_{ij}^{00} + Y_{ij}^{11}) - (r_{ij}^C + 1)/2|) \\ &= \textstyle\sum_{1 \le i < j \le N} (r_{ij}^C - r_{ij}^R) \cdot (Y_{ij}^{00} + Y_{ij}^{11} - Y_{ij}^{01} - Y_{ij}^{10}) \end{aligned} \tag{5}$$

It is worth pointing out that because the signed allele correlation r ranges from -1 to 1, for the validity of the test statistic, we compare the pair-wise allele frequency of the victim (Y_{ij}) with $(r+1)/2$ of the case and reference groups, respectively, which are non-negative (ranging from 0 to 1). Based on the null hypothesis that the to-be-tested sequence (of the victim) is not within the case group[4], we have $E(T_r) = 0$. However, if the sequence is indeed in the case group, the expected contribution of this instance to a specific signed allele correlation r_{ij}^C in the summation $\frac{\partial T_{ij}}{\partial C_{h_i h_j}}$ is non-negative for any pair of SNPs i and j and any pair of alleles 11, 10, 01 or 00 (see Appendix 1 for a proof). Therefore, the sum statistic of T_r is valid. In Appendix 2, we also give a sketch of the proof that T_r is close to an optimal test statistic assuming r approximately follows a normal distribution. Notably, even though the test statistics proposed here has a similar form as the one proposed by Homer et al. [42], it sums over $\binom{m}{2}$ variables (i.e. signed allele correlations) instead of m independent SNPs, and hence, as we will show in Section 4.2, it results in a much more powerful attack.

Markov model estimation. However, since the signed allele correlations (r_{ij}) are not completely independent, we cannot assume the distribution of T_r under the null hypothesis is normal. With the limited size of available reference population from the HapMap project, we resorted to the Markov chain modeling and sampling techniques to simulate data for estimating the confidence of our statistical attack. Markov models, both (inhomogeneous) Markov Model [38, 47, 52] and Hidden Markov Model [46, 52], have been used extensively in modeling SNP sequences. In our research, we constructed an inhomogeneous 1st order Markov model by a maximum likelihood approach from a limited number of haplotypes obtained from HapMap website (as the training dataset). The model parameters, including one set of initial probabilities and $m-1$ set of transition probabilities, can be directly estimated based on the counts for single SNPs and the neighboring SNP pairs in the training dataset. Once the Markov model is built, it can be used to draw unlimited number of haplotypes with a similar genetic background as the ones in the training dataset. In each simulation, at least 1000 case and reference groups (each with 50-1000 haplotypes) are sampled, and the test statistics T_r is computed for each of these paired case/reference groups. Using this technology, we studies how the power of the test statistics is affected by various GWAS aspects such as the SNP sequence of the victim, the case group size, the resemblance between the reference and case group, the precision of reported r^2 and missing values of r^2, etc.

[4]Note that this is sufficient for determining one's presence in the case, because common sequences can actually be present in both the case and reference populations.

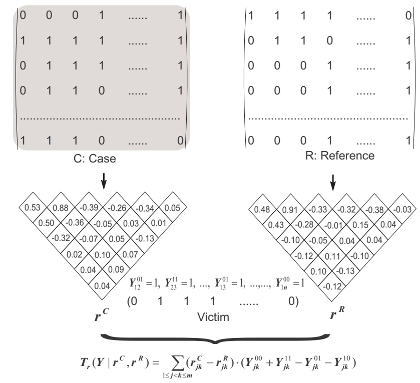

Figure 2: A schematic illustration of a statistical attack on the reported r^2 values between pairs of SNPs. Assume the attacker has obtained the haplotype of a victim and his goal is to determine how likely this victim is within a case group in a GWAS study. To achieve this goal, the attacker will compute a test statistics (T_r) that measures the distance between two set of *signed* allele correlations, r^C derived from the reported (r^2) values computed from the haplotypes of the individuals in the case group (left panel), and r^R computed directly from haplotypes of a mixture of the victim and the reference group (right panel). T_c has the same property as the test statistics D proposed by Homer et al.: it is significantly larger when the victim is within the case group than otherwise. Using an empirically deduced distribution of T_c given the null hypothesis that the victim is not in the case group, the attacker can then estimate the confidence level of identifying the victim.

Reference population. A practical attack needs a reference population that preferably resembles the case group in the distribution of r. Such a population can come from various sources, for example, replication studies on different populations, or reuse of the case data for a new study. Among these, of a particular interest is the possibility that one could actually acquire the genomes of the reference population through signing an agreement. Though privacy agreements legally bind researchers against revealing the genomic data under study, it does nothing to stop an attacker using the collected data as a reference population to compromise the privacy of individuals involved in other datasets, for which she has no obligation. Actually, in Section 4.2, we demonstrate that even a bad reference, with an r distribution deviated from the case group, can allow us to identify some individuals with high confidence.

Encoding nuclotides. A problem an attacker has to tackle is to translate the victim?s nucleotides (A, T, C, G) acquired from genotyping into alleles (0 or 1). These alleles are determined from the population studied in a GWAS: for the two nucleotide values a SNP can take, the popular one is encoded as 0, and the other as 1. The attacker, who does not have direct access to the population, needs to figure out how to encode the victim's nucleotides before launching the attack. A solution is to utilize the HapMap to identify major alleles, which is often very effective, as observed in our research.

3.3 An Integer-Programming Attack

Given pair-wise allele frequencies for a whole locus that involves tens of or even a couple of hundred SNPs, the idea of recovering individuals' SNPs can become really enticing. We believe that this can actually be done with proper techniques. In this section, we report such an attack based upon integer programming.

Illustrated in Figure 3 is the sequences of N individuals, each containing M SNPs. These SNPs form an N by M matrix, below which are their LDs. Our attack attempts to recover the whole or part of the matrix from the LDs that are in the form of pair-wise allele frequencies. To this end, we designed a "divide-and-conquer" approach described below.

ID	Sequence
1	0 1 1 0 1 1 1 0 0 1 0 1 0 0 1 1 1 0 1 … … … … … 1 0 0 1 1 1 0 1 1 1 1 0 0 0 1 1 1 0 0 1 0
2	1 1 1 0 1 0 1 0 1 0 0 0 1 1 0 0 1 1 0 … … … … … 0 0 0 0 0 0 0 1 1 0 1 0 0 1 0 0 0 0 0 0 1
3	0 1 1 1 0 0 0 1 0 1 0 0 1 1 1 0 1 0 1 … … … … … 0 0 1 1 1 0 0 0 1 0 1 0 1 1 0 0 0 1 1 1 1
…	… … …
…	… … …
n-1	0 0 1 0 1 1 1 1 0 0 0 0 0 1 1 0 0 1 1 … … … … … 0 1 1 0 1 0 1 0 1 1 0 0 1 0 1 0 0 1 1 0 1
n	1 0 0 1 0 1 0 0 1 0 1 0 1 1 0 1 0 1 0 … … … … … 1 0 1 1 0 0 0 0 1 1 1 1 0 0 1 0 0 1 1 1 1

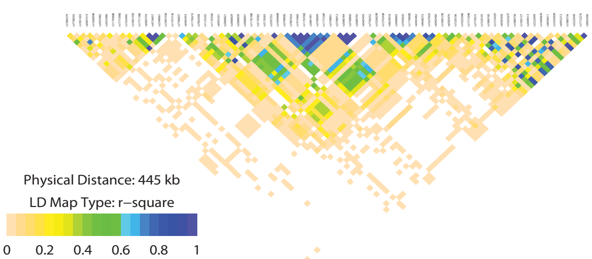

Figure 3: Recover individuals' SNPs

Human DNA information is passed from one generation to another in a way that recombinations occur much more frequently between two SNP blocks, called *haplotype block*, than within a block. As a result, a haplotype block typically contains only a small number of combinations (i.e., haplotypes) of SNP values. To leverage this property, our attack partitions a SNP sequence into blocks according to their LDs, to ensure that the SNPs on the same block have strong connections. Then, integer programming is utilized to find out the haplotypes within individual blocks that satisfy the constraints of pair-wise frequencies and single-SNP frequencies. Finally, haplotypes in different blocks are connected based upon the LDs between these blocks.

Recovering individual blocks. GWAS participants' SNP sequences within a block are actually described by the counts of different haplotypes. Therefore, the first stage of our attack focuses on determining those counts. Let $[x_0 \cdots x_{2^l-1}]$ be a vector of integer variables that represent the counts of individual haplotypes, where l is the number of the SNPs on a block. Note that individual SNP can only have two values, and thus 2^l is the total number of different haplotypes in the block. To find a solution for the vector, our approach takes the following three steps:

• Step 1: We first build a system of linear equations upon the equality constraints derived from pair-wise allele counts and the number of the participants. Figure 4 describes an example that works on a 3-SNP block. The block can have total 8 possible haplotypes whose counts are denoted by $\vec{X} = [x_0 \cdots x_7]$. Given 12 pair-wise frequencies (each pair can take 4 alleles), we can build 12 linear equations: the first equation represents the count of the allele "00" for the first two SNPs, the second is the count of "01" for the same SNP pair, and other pair-wise counts are described by the remaining equations. Such a system is then solved using Gaussian elimination, and its solution set can be represented in a parametric form that includes a set of free variables. In the example, the solutions are described by $\vec{X} = \vec{a}x_7 + \vec{b}$, where $\vec{a}$ and $\vec{b}$ are two constant vectors as illustrated in Figure 4, and x_7 is a free variable.

•Step 2: The ranges of the free variables are determined by integer programming. We first add in inequality constraints that require all the solutions to be non-negative. Then, for every free variable, a pair of integer programming problems are solved to minimize and maximize its value under the constraints. This gives us the acceptable range of the variable. Figure 4 displays the inequality constraints for the example, under which $\min x_7$ and $\max x_7$ reveal that the variable can take 0 or 1.

•Step 3: In the presence of multiple free variables, we need to exhaustively search their value ranges to find a combination that satisfies all the constraints. This is an exponential problem. However, its scale can be controlled by adjusting block size. Every combination found is fed into the parametric solution of the linear equation system, which gives us the haplotype counts we are looking for.

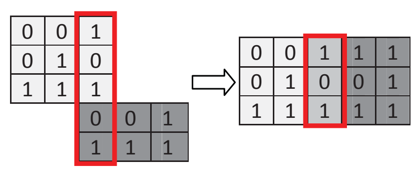

Figure 5: Connect two blocks through overlapping.

Connecting different blocks. After restoring haplotypes for individual blocks, we move on to link different blocks together based on the LDs between them. The SNPs in different blocks typically have weak correlations. However, the aggregate connections from individual SNP pairs can be strong between two large blocks. We took a strategy in our research that first merges blocks with strong LDs into a large block, and then bridges it to other blocks. To connect two blocks, we again use a vector of variables to represent the counts for individual combinations of the haplotypes from different blocks. For example, consider two 3-SNP blocks, each with two haplotypes: ("001" and "011") for one and ("100" and "110") for the other; our approach generates 4 variables to represent the counts for ("001" "100"), ("001" "110"), ("011" "100") and ("011" "110") respectively. In general, two blocks with m and n different haplotypes respectively bring in mn variables. The solutions for those variables are computed through integer programming, as does the step for recovering individual blocks.

A problem arises when the number of haplotypes in each block is large, which makes the number of variables even larger. As a result, the time integer programming takes to find a solution can increase exponentially. In our research, we adopted a simple technique to mitigate this problem. For the part of a SNP sequence where LDs are weak, our approach cuts blocks in a way that allows two blocks to share a set of SNPs. Those SNPs let us look at the common part of the haplotypes from different blocks, and as a result, help

Sequence		
0	0	0
1	0	1
1	1	0
0	1	1
1	1	0

Input	
x_0	000
x_1	001
x_2	010
x_3	011
x_4	100
x_5	101
x_6	110
x_7	111

Step 1

$$\begin{cases} x_0 + x_1 = 1 \\ x_2 + x_3 = 1 \\ x_4 + x_5 = 1 \\ x_6 + x_7 = 2 \\ x_0 + x_4 = 1 \\ x_1 + x_5 = 1 \\ x_2 + x_6 = 2 \\ x_3 + x_7 = 1 \\ x_0 + x_2 = 1 \\ x_1 + x_3 = 1 \\ x_4 + x_6 = 2 \\ x_5 + x_7 = 1 \end{cases} \Rightarrow \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \overset{\vec{a}}{\begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}} x_7 + \overset{\vec{b}}{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 2 \\ 0 \end{pmatrix}}$$

Step 2

$$\begin{matrix} min\ x_7 \\ \Rightarrow \quad x_7 \in [0,1] \\ max\ x_7 \end{matrix}$$

$$\begin{pmatrix} x_0 \geq 0 \\ x_1 \geq 0 \\ x_2 \geq 0 \\ x_3 \geq 0 \\ x_4 \geq 0 \\ x_5 \geq 0 \\ x_6 \geq 0 \\ x_7 \geq 0 \end{pmatrix} \Rightarrow \begin{pmatrix} -x_7 + 1 \geq 0 \\ x_7 \geq 0 \\ x_7 \geq 0 \\ -x_7 + 1 \geq 0 \\ x_7 \geq 0 \\ -x_7 + 1 \geq 0 \\ x_7 + 2 \geq 0 \\ x_7 \geq 0 \end{pmatrix}$$

Step 3

$$\begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{pmatrix} = \begin{cases} if\ x_7 = 0, solution = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} \\ if\ x_7 = 1, solution = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \end{cases}$$

Figure 4: Recover individual blocks

reduce the number of variables. Consider two blocks, A and B, with an overlap that includes k haplotypes. For each haplotype i ($1 \leq i \leq k$), suppose that it is attached to m_i different haplotypes in A and n_i in B. This gives us at most $\sum_k m_i n_i$ different ways to connect the haplotypes from different blocks. Figure 5 presents an example in which two blocks with 3 and 2 haplotypes respectively have an overlap involving 2 haplotypes. We need 3 variables to describe all possible haplotype combinations between these blocks, instead of 6.

Identification of an individual. Once we obtain a solution (a set of SNP sequences), we could compute the corresponding test statistic T_r (based on the r values for the case, and a reference sample available) and estimate the identification confidence of each of the sequences. If one sequence receives a small p-value, it is probably a correct solution, and the individual with the sequence (known to us now) is probably in the case group; on the other hand if a sequence receives a large p-value, the sequence could either be wrong or not unique enough to determine a person. Actually, a sequence with a high identification confidence can reveal a lot of information about its owner. For example, if it happens to contain the SNPs related to observable phenotypes, Malin's attack [49] could be applied to identify the individual associated with the sequence, even if the attacker does not have the victim's DNA profile *a priori*.

3.4 Limitations of Simple Defense

Immediate mitigation of the threat to test statistics includes reducing the precisions of published statistics, or publishing less data. These measures will undoubtedly make our attacks more difficult to succeed. However, their practical implementation is by no means trivial: without a well-thought-out plan, they will either significantly undermine the scientific value of GWAS papers, or fail to stop our attacks. In this section, we discuss the impacts those measures can have on our techniques.

Low-precision statistics. As described in Section 3.1, our approach recovers allele frequencies by propagating genetic markers' frequencies to other SNPs through their LDs. The markers' frequencies are among the most important outcomes of a GWAS and therefore have to be released. What the defender can do here is to downgrade the precisions of LD statistics. This, unfortunately, is often insufficient for blocking the information that can be used to recover allele frequencies. The fundamental problem here is that all the statistics, p-values of individuals SNP and r^2 for SNP pairs, are correlated. Such correlations, together with the marker's frequencies and the size of a case group, can make up for the information loss caused by coarse-grained statistics. Consider the example in Figure 1: the LD between SNPs (S_1, S_2), affects the relation between (S_2, S_3), given the constraints of the total number of participants; the LDs of (S_1, S_2) and (S_2, S_3) further constrain that of (S_1, S_3). Our approach leverages such relations and therefore is very robust to inaccurate statistics. An experimental study reported in Section 4 shows that given r^2 rounded to 2 decimal places (only 1 decimal place for r), we still restored more than 50% of pair-wise allele frequencies and all the signs.

On the other hand, the attacks we propose do not rely on fine-grained statistics. Our statistical attack only needs r rounded to 1 decimal place to establish one's presence in the case (Section 4.2). The integer-programming attack can still work in the absence of some frequency constraints, though this can result in multiple solutions. These properties of our attacks were evaluated through an experimental study, which is reported in Section 4.

Thresholds. Publishing less data can certainly make it more difficult for an attacker to infer sensitive information. However, it equally renders GWA papers less informative. An obvious solution is to use a threshold to remove the data deemed insignificant to the research. The question is how to set such a threshold. For example, in Figure 3, all the r^2 values below 0.01 are dropped. However, we can still figure out haplotype frequencies for individual haplotype blocks using the r^2 within blocks, and connect different blocks together by running a maximum likelihood estimator over the remaining LDs between blocks. Moreover, if the remaining data contains sufficient information for recovering signs, our statistical attack still works.

We believe that a one-size-fits-all threshold does not work for GWAS. Techniques need to be developed to assess the outcomes of individual research to compute a customized threshold that enables dissemination of the findings of a study without compromising the privacy of its participants.

4. EVALUATIONS

This section reports an experimental study of the techniques we propose. The objective is to demonstrate that the threats of our attacks are realistic. Like the prior work [42], our study was based upon the real haplotypes from the HapMap project [44] (`http://www.hapmap.org/`). More specifically, we used haplotype sequences at FGFR2 locus (around SNP rs1219648) in the HapMap phase 3 release 2, which covers 200kb region from SNP rs12354864 (human reference genome b36 location 123189345) to SNP rs7900009 (human reference genome b36 location 123450068). The locus was chosen because it was recently reported in a GWAS paper [43] to be associated with the risk of sporadic postmenopausal breast cancer, and the linkage disequilibrium plot with measures of r^2 for the 174 SNPs in the locus is directly available in the paper (as did routinely in most GWAS articles), which makes it a proper target of our attack. Note the real haplotypes used in our evaluation are *not* from the individuals used in the study reported by the GWAS paper [43],

Table 3: Infer Frequencies and Signs

Statistics Precision		Recovered Information %		
r^2	p-value	single SNP frequency	pair-wise frequency	sign of r
0.1	0.1	12.1	1.8	6.7
0.1	0.00001	40.6	11.7	31.7
0.01	*	100	50.1	98.7
0.001	*	100	90.4	100
0.0001	*	100	95.1	100

but instead from the HapMap database, due to privacy concerns[5].

In our experiment, we randomly chose a population of 200 with Africa backgrounds, half as cases and other half as controls. Association statistics of the population were first computed, including pair-wise r^2 of the case group and p-values of individual SNPs estimated from Pearson?s chi-square. Our attacks were evaluated over these statistics, together with the knowledge of the sizes of the case and control groups and the allele frequencies of a SNP treated as a genetic marker (SNP rs1219648). Such information is typically reported in a GWAS paper and its supplementary materials, or at the very least, can be conveniently acquired from authors of the paper. Following we describe our evaluation of the techniques for inferring allele frequencies, identifying individuals and recovering SNP sequences.

4.1 Inferring Allele Frequencies and Signs

The first step of our attack was to infer the allele frequencies for both individual SNPs and SNP pairs from the statistics. Our approach started with the marker and propagated its frequencies to other SNPs through their r^2. During this process, a constraint solver (cream [1]) was used to search for the unique solutions for the allele counts and the pair-wise allele counts within the population under the constraints posed by the r^2, the p-values and the population sizes, as described in Section 3.1. The solver also recovered the signs of r, if they could be uniquely determined. In our experiment, we evaluated this approach against the statistics with various precisions. The outcomes are illustrated in Table 3.

The experiment demonstrates the prowess of our technique. Given an r^2 rounded to 2 decimal places, which amounts to 1 decimal place for r, we successfully recovered all the single-SNP frequencies, half of pair-wise frequencies and almost all the signs for r, regardless of the precisions of p-values. 95% of pair-wise frequencies were identified when the precision of r^2 reached 4 decimal places. An interesting observation is that p-values could make up for the loss of precision in r^2: an r^2 of 1 decimal place, once paired with a p-value of 5 decimal places, could be used to restore more than 30% of signs. On the other hand, GWAS papers typically offer p-values rounded to 4 to 5 decimal places (often in the supplementary materials [58, 54, 64, 40]), and r^2 even to 8 decimal places [43]. This is more than enough for our attack.

4.2 Identifying Individuals

Evaluation of the power of the T_r statistic based on simulation experiments. Based on reference groups drawn from the first order inhomogeneous Markov chain (see section 3.2), we estimated the power of our statistical attack. If not mentioned otherwise, the simulation model is built from the FGFR2 loci with 230 SNP sequences of the YRI population (Yoruba in Ibadan, Nigeria) from the HapMap database. We first compared our attack with the one proposed by Homer et al. We found that the power of statistical attack on r^2 using the T_r statistic is much higher than the attack on single SNP profiles, using the statistic $T_p = \sum_j D(Y_j)$, where $D(Y_j)$ is defined in equation 1, akin to the statistc proposed by Homer et al. (see Figure 6). Let H_0 be the null hypothesis (the victim is not in the case group), and let H_A be the alternative hypothesis (the victim is in the case group). We note that the power of the statistic is higher if the overlapping portion of the two distributions (under H_0 and H_A, respectively) is smaller (Figure 6). We estimate the distribution of H_0 by sampling the victim and the cases independently, and the distribution of H_A by including the victim's SNP sequence into the sampled case group. For 200 sequences of 174 SNPs at the FGFR2 locus in each of the case and control groups, T_r statistic can identify 80% individuals in the case group at fixed type I error of 0.05 (or 95% confidence), while T_p statistic can only identify around 9%. Note that the lower bound of the power for any statistic equals type I error (5%); thus T_r statistic is about 20 times more powerful. For T_p to reach a similar power, around 30 times more SNPs are required (i.e. 5000 SNPs).

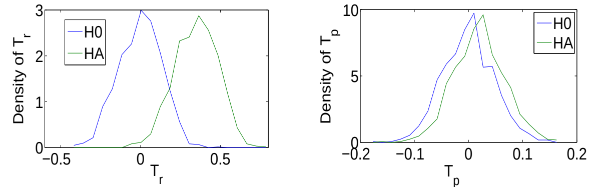

Figure 6: Comparison of estimated probability density functions of the test statistic on pairwise r values (T_r, left) vs. the test statistic based on single SNP frequencies (T_p, right). The probability density is estimated based on 1000 rounds of simulations with 1 fixed victim, 200 cases and 200 references.

We next analyzed the performance of the statistical attack under the setting of different parameters, including the SNP sequence of the victim, the case group size, quality of the reference group and the precision of reported r^2. Using the simulation based on the Markov model for human populations, we found that T_r approximately follows a normal distribution (Figure 6), and the *standard deviation* of the distribution under the null hypothesis is nearly constant with respect to various SNP sequence of the victim and varous ethnic group the cases belong to, as long as the sample size and number of SNPs are fixed. This implies that we may not need to re-deduce the null distribution for every test carried out. Interestingly, on the contrary, the power of attack actually depends on all these factors. In a set of simulations with 100 SNPs, 200 cases and 200 references, we found that power of attack can vary from 30% to 80% (with average 62% and standard error 11%) depending on the SNP sequences of the cases. When we looked into the ethnic group of the cases, we found that the power for cases from Africa (population YRI, Yoruba in Ibadan, Nigeria) is higher than those from central Europe or east Asia (data not shown), indicating that some individuals (from some ethnic groups) are easier to be identified than others, presumably because they carried more sensitive SNPs than the others. Finally, we examined the power of attack on cases with various sizes, ranging from 50 to 1600 (table 4). Note that even when the case is very large (e.g. 1600, larger than typical GWAS studies), there are still a significant proportion (18.1%) of cases who can be identified confidently. This result suggested that with a single locus (of 174 SNPs), our statistical attack has the potential to identify many individuals in a typical GWAS study.

We evaluated our attack against two types of defenses, namely the *low precision* approach (only providing low precision values for r in the GWAS paper) and the *threshold* defense (removing r values below the threshold). We found the performance of the statistical test is very robust to both countermeasures. At very strong defense level, e.g. threshold 0.1 for $|r|$, or the precision level of

[5] Actually, to acquire the genome data from a real GWAS, one needs to sign an agreement that forbids use of the data for personal identification.

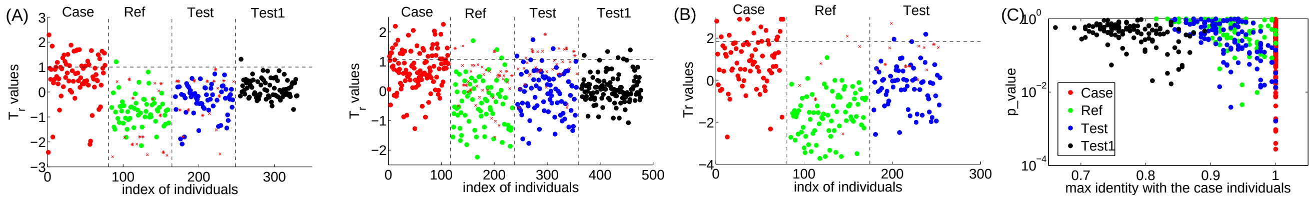

Figure 7: Evaluation of the power of the statistical attack. (A) Realistic attacks using good references. SNP sequences retrieved from Hapmap database are used as cases (*Case*, red dots), references(*Ref*, green dots) as well as the tests (*Test*, blue dots and *Test1*, black dots). Individuals in *Ref*/*Test*/*Test1* are marked if they have 100% (*) or $> 95\%$ (x) identity with some individuals in *Cases*. In the left plot, *Case*/*Ref*/*Test* SNP sequences are taken from *YRI* (Yoruba in Ibadan, Nigeria), while *Test1* haplotypes are from a different population *JPT+CHB*(Japanese in Tokyo, Japan, and Han Chinese in Beijing, China). In the right plot, *Case*/*Ref*/*Test* are from *JPT+CHB*, with *Test1* from *YRI*. (B). Realistic attacks using average references. Legend should be interpreted the same as in *A* except that there is no *Test1*. Here cases are taken from *yri* population, whereas references are taken from a different (but related) population *ASW* (African ancestry in Southwest USA) (C). Estimated p-values for the attack shown in the right plot of *A*. The variance under the null distribution were estimated by T_r values from the groups of *Test* and *Test1*.

Table 4: Dependence of the statistical power of the T_r on sample size based on the simulation on the FGFR2 locus. N is the number of SNP sequences in the case groups (the same as the number in the reference group).

N	50	100	200	400	800	1600
power (%)	99.9	85.7	67.2	40.4	36.2	18.1

Table 5: Percentage of statistical power (at 0.05 Type I error) left at various precision of input data r^C. The power is estimated based on 1000 rounds of simulated attacks. The number of cases and controls are both 200. The individual SNP sequences were randomly draanw from the inhomogenous Markov Chain built on 230 SNP sequences of the FGFR2 locus from the HapMap phase 3 YRI (Africa) population.

Precision of r^C	0.5	0.2	0.1	0.01	0.001
% power π left	12	74	85	100	100

1 decimal place, majority of the statistical power (79% and 85% respectively) is retained. The results are shown in table 5.

The power of realistic attacks. The simulations described above can generate a large set of SNP sequences, thus enabled us to systematically analyze the power of our statistical attack. As a further step, we want to evaluate the performance of the statistical attack with a realistic setting. Three types of attacks were experimented (Figure 7). In each of them, three groups (C - case group, R - reference group, and T - test group) of real SNP sequences of FGFR2 locus are retrieved from different populations in HapMap database (see section 3.2 for details); then, for each SNP sequence H from $Case$, $Reference$, and $Test$, T_r is computed based on H and r values from $Case$ and $Reference$ (see figure 2). Note that the individuals in the test group are completely independent from the case and the reference, and therefore the mean of their T_r is zero. In the first two types of attacks, $Case$ and $Reference$ individuals are taken from a same population (YRI, Yoruba in Ibadan, Nigeria, or JPT+CHB, Japanese in Tokyo, Japan, and Han Chinese in Beijing, China). This mimics the situation where we have a good reference sample that resemble the case group very well. In the third type of experiments, $Case$ (from YRI) and $Reference$ (from ASW - African ancestry in Southwest USA) are taken from a different but related population. This mimics the situation where we do not have a good reference sample, but have to use an *average* reference sample available from a public data source. For each experiment, with the case and reference SNP sequences fixed, T_r value is computed for each SNP sequence in the case, reference, and the test groups.

The results from these realistic attacks are promising. For the two experiments with good references, many individuals in the case group receiving higher T_r values than individuals from other groups (Figure 7B), suggesting that many SNP sequences in the case group can indeed be identified with relatively high confidence, with p-values from 0.01 to 10^{-5} (Figure 7C and data not shown). While for the situation where an average reference is used, there will be a drop of the discrimination power (Figure 7C compared with figure 7B right), however there is still a significant amount of power left, making the attack still possible. Note that in many GWAS studies, data for more than one loci are provided, which makes the attack is even more powerful.

To estimate the confidence (p-values) for each victim, we only need to estimated the variance of T_r under the null hypothesis [6]. One approach is to build a Markov chain from the reference sample (which is known to the attacker), and then estimate the variance by the simulation experiment (see 3.2 for details). A second approach is to estimate the variance directly from T_r values obtained from individuals of the test sample. It turns out the second approach gives relatively more conservative estimations, whereas the first approach gives over-optimistic estimation that may lead to a lot of false positives. Figure 7C gives an example of estimated p-values for the attack with a good reference. One interesting observation is that, many non-case individuals could also receive relatively small p-values, especially when they are not completely independent from some case individuals (Figure 7C). This is expected, since the objective of our statistical attack, essentially, is to test the (genetic) independence of the victim with the case individuals.

4.3 Recovering SNPs

We implemented the Integer-Programming attack using Matlab, based upon two toolboxes, $rref$ [6] for solving systems of linear equations, and $bintprog$ [3] for integer programming. This implementation was run on our dataset to recover the SNP sequences for all 100 individuals from their pair-wise allele frequencies. In the attack, we partitioned the sequences into small blocks according to the LDs of the locus, as demonstrated by the r^2. Block size varied around 10 SNPs. Haplotypes within individual blocks were first restored through solving linear-equation systems and integer programming, and then connected together using overlapping blocks in accordance to the LDs, as described in Section 3.3. This attack was run on a system with 2.80GHz Core 2 Duo processor and 3GB memory. Within 12 hours, we successfully restored the 174 SNPs for all 100 participants. This demonstrates that the threat of the integer programming attack is realistic.

5. DISCUSSION

[6]We may need to estimate the expected values of T_r if an average reference sample is used.

A GWAS can either analyze individuals' genotypes or phase them into haplotypes before the analysis happens. The maturity of phasing technologies makes the use of haplotypes, which carries more information, increasingly a trend [62, 60, 20, 15, 26, 29]. Although we present our attacks in the context of haplotypes, some of the techniques involved can actually be applied to genotypes. Specifically, our statistical attack can still identify an individual from pair-wise allele frequencies estimated from genotypes, as long as such estimate does not switch the signs of r^C in a large scale. The integer programming attack may not be directly applicable to genotypes, because it needs consistent pair-wise frequencies. However, we could use it to restore genotypes if pair-wise frequencies reported by GWAS papers can be mapped back to the frequencies of genotype. This is possible if we can identify the inputs of the MLE that builds those frequencies from its outputs. Further exploration of this direction is left to future research.

The attacks we propose could be defeated by well-planned countermeasures. A potential approach can be adding noise to published dataset. This, however, needs to be done carefully, because the noise can undermine the scientific values of a paper, making others hard to repeat the experiments reported by a GWAS. One technique we can use is to adjust some of the published LDs to the extent that the signs of r are just changed. This requires a careful selection of a set of SNP pairs, whose LDs are not significantly affected by switch of signs. Examples of the SNPs are those with very close $C_{00}C_{11}$ and $C_{01}C_{10}$. Also important here is the assurance that other unperturbed LDs do not give an attacker sufficient statistical powers to identify an individual. Another approach is to selectively remove some data. For example, dropping some LD statistics can interrupt the path for propagating markers' frequencies, and as a result, makes allele frequencies more difficult to recover. Fundamentally, effective defense against our attacks can be achieved by enforcing differential privacy: that is, the presence of an individual in the case group changes nothing but negligible statistical features of the group. For example, GWAS researchers can use our statistic to check the data to be published: if an individual's SNP sequence, once added, noticeably alters r, they can decide to remove it from the dataset. We will study such defense techniques in the follow-up research.

6. RELATED WORK

The problem of releasing aggregate data in privacy-preserving ways has been intensively studied in the areas like privacy-preserving data analysis [31, 36], statistical disclosure control [19, 21, 34], inference control [25], privacy-preserving data mining [12, 13], and private data analysis [32, 53].

Privacy problems in GWAS, however, are related to special structures of genome data (linkage disequilibrium) and background information (reference populations), which have not been studied in prior research. Recently, research has been conducted on privacy-preserving genome computing [17, 45, 23]. Those approaches are more to do with preventing a party from accessing sensitive data, than protecting sensitive information from being inferred from the outcome of a computation. A recently proposed concept highly related to our attacks is *differential privacy* [31]. Loosely speaking, the concept ensures that removing or adding an individual's record to a database does not substantially changes the statistics calculated from the database. So far, this has been achieved through adding noise [18, 32, 21]. Such an approach, if carefully designed, could mitigate the threat of the attacks we propose. For example, if the noise is used in a way that only the signs of some LDs are changed, the statistical power of our attack can be weakened while the scientific values of the paper can be largely preserved. Research on this direction is left as our future work.

Recovering individuals' SNP sequences is essentially the problem of contingency table release, which has been studied in statistics community for decades [22, 41, 24, 63, 27] in the context of census. Consider an individual's record in a database as a row, which consists of k binary attributes. A contingency table is a vector that describes, for each combination of k attributes (called a *setting*), the counts of the rows satisfying this setting. Typically, only the counts (called *marginals*) of different settings are published, instead of the table. What an attacker wants to do is to infer the table from the marginals.

In our integer programming attack, individual SNP frequencies and pair-wise frequencies are actually such marginals. However, existing techniques [63] cannot be directly applied to recover SNPs from them, because of the scale of our problem: up to our knowledge, prior research can only handle as many as 16 attributes [27], while our attack needs to work on 174 attributes. This was achieved in our research by taking advantage of special properties of genomic data. Our "divide-and-conquer" approach made full use of the correlations among different SNPs, which can be observed from published r^2. More specifically, we first partitioned a SNP sequence into small blocks according to their LDs, which significantly increases the chance for restoring these blocks because the LDs within the same blocks are strong; then, individual blocks were connected using their aggregate correlations. The techniques designed for this purpose, solving linear equations and use of overlapping blocks in particular, are novel, up to our knowledge.

GWAS, with its' first papers came out in 2005 [33, 39], is a new population genetics approach that was enabled by the high throughput genome technology and advanced computational methods [51]. As a technique to discover previously hard-to-find disease susceptibility genes, it has shown great potentials in promoting the understanding of human diseases [51]. While GWAS methods continues being improved thanking to the efforts of many researchers, few researcher has looked into the privacy risks in GWAS [48, 37]. The most relevant work other than Homer et al. is the association study based on the *pooled* genotypes, i.e. the aggregate genotype profiles of a group of cases and controls (rather than for each individual of them) were used in the SNP marker discovery, e.g by Yang et.al. [61]. These approaches, though adopting a similar experimental setting, aim to address a distinct problem as our approach, that is, to retrieve SNPs (or SNP sequences) significantly associated with the disease, which tend to be the common genetic features of the disease population rather than the features specific to the individuals in the case group, and hence cannot be used for the individual identification.

7. CONCLUSION

GWAS is among the most active research areas in biomedical research. It is also the area fraught with privacy concerns. The recent work by Homer et al. [42] demonstrates that personal identification is plausible by analyzing a large number of allele frequencies related to GWAS. The privacy threat of this kind has been found in our research to be even more realistic than expected. In this paper, we report two new techniques that can lead to identification of the participants of a GWAS from a small set of statistics, as those routinely published in GWAS papers. One of the techniques can statistically determine the presence of an individual in the case group, based upon the LDs among as few as a couple of hundred SNPs. The other attack can even recover all participants' SNP sequences related to the statistics. We also show that these attacks work on coarse-grained statistics. Our experimental study further justifies the concerns of such threats, which were shown to be capable of cracking statistics computed from real genome data.

A further step in this important direction will be evaluation of information leaks in GWAS and development of effective countermeasures. The tendency of sharing and reusing the data from prior GWAS will certainly open new avenues for attacks. On the defense side, we believe that research in statistical disclosure control, differential privacy in particular, can offer an effective guideline to mitigate and ultimately eliminate the privacy threat to GWAS.

8. REFERENCES

[1] Cream: Class library for constraint programming in java. `http://bach.istc.kobe-u.ac.jp/cream/`, 2003.

[2] Haplotype estimation and association. `http://slack.ser.man.ac.uk/theory/association_hap.html`, 2005.

[3] bintprog: solve binary integer programming problems. `http://www.mathworks.com/access/helpdesk/help/toolbox/optim/index.html?/access/helpdesk/help/toolbox/optim/ug/bintprog.html`, 2008.

[4] Company to sequence 1,000 human genomes in 2009 for $5,000 each. `http://www.medicalnewstoday.com/articles/124572.php`, 2008.

[5] Nih background fact sheet on gwas policy update. `http://grants.nih.gov/grants/gwas/background_fact_sheet_20080828.pdf`, 2008.

[6] rref: Reduced row echelon form. `http://www.mathworks.com/access/helpdesk/help/techdoc/index.html?/access/helpdesk/help/techdoc/ref/rref.html`, 2008.

[7] Genome-wide association studies. `http://grants.nih.gov/grants/gwas/`, 2009.

[8] International hapmap project. `http://www.hapmap.org/`, 2009.

[9] The origins of haplotypes. `http://www.hapmap.org/originhaplotype.html.en`, 2009.

[10] Policy for sharing of data obtained in nih supported or conducted genome-wide association studies (gwas). `http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html#principles`, 2009.

[11] The r project for statistical computing. `http://www.r-project.org/`, 2009.

[12] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255, New York, NY, USA, 2001. ACM.

[13] R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000.

[14] A. Agresti. *Categorical Data Analysis*. New York: Wiley-Interscience, 2002.

[15] C. I. Amos et al. Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1. 40(5):616–622, 2008.

[16] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(4):375–386, 1955.

[17] M. J. Atallah, F. Kerschbaum, and W. Du. Secure and private sequence comparisons. In *WPES '03: Proceedings of the 2003 ACM workshop on Privacy in the electronic society*, pages 39–44, New York, NY, USA, 2003. ACM.

[18] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS '07: Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 273–282, New York, NY, USA, 2007. ACM.

[19] L. L. Beck. A security machanism for statistical database. *ACM Trans. Database Syst.*, 5(3):316–3338, 1980.

[20] T. Becker and C. Herold. Joint analysis of tightly linked snps in screening step of genome-wide association studies leads to increased power. 2009.

[21] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS '05: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, New York, NY, USA, 2005. ACM.

[22] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilitÃă. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8(1):3–62, 1936.

[23] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls. Privacy-preserving matching of dna profiles. Technical Report Report 2008/203, ACR Cryptology ePrint Archive, 2008.

[24] Y. Chen, P. Diaconis, S. P. Holmes, and J. S. Liu. Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100:109–120, 2003.

[25] F. Y. Chin and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Trans. Softw. Eng.*, 8(6):574–582, 1982.

[26] T. DA et al. Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nature Genetics*, 41(3):283–285, February 2009.

[27] A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables induced by fixed marginal totals. *Statistical Journal of the United Nations ECE*, 18:363–371, 2001.

[28] J. Domingo-ferrer. Advances in inference control in statistical databases: An overview. In *In Inference Control in Statistical Databases: From Theory to Practise (2002*, pages 1–7. Springer Verlag, 2002.

[29] A. Doring et al. Slc2a9 influences uric acid concentrations with pronounced sex-specific effects. 40(4):430–436, 2008.

[30] R. H. H. Duerr et al. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, October 2006.

[31] C. Dwork. Differential privacy. In *in ICALP*, pages 1–12. Springer, 2006.

[32] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 2006.

[33] A. O. Edwards, R. Ritter, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer. Complement factor h polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, Apr 2005.

[34] S. E. Fienberg. Datamining and disclosure limitation for categorical statistical databases. In *Proceedings of Workshop on Privacy and Security Aspects of Data Mining, Fourth IEEE International Conference on Data Mining (ICDM 2004*, pages 1–12. Nova Science Publishing, 2004.

[35] R. Fisher. On the interpretation of chi-square from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[36] J. Gehrke. Models and methods for privacy-preserving data analysis and publishing. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*, page 105, Washington, DC, USA, 2006. IEEE Computer Society.

[37] H. T. Greely. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. *Annu Rev Genomics Hum Genet*, 8:343–364, 2007.

[38] G. Greenspan and D. Geiger. Modeling haplotype block variation using markov chains. *Genetics*, 172(4):2583–2599, Apr 2006.

[39] J. L. Haines, M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. A. Postel, and M. A. Pericak-Vance. Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, Apr 2005.

[40] H. Hakonarson et al. A genome-wide association study identifies kiaa0350 as a type 1 diabetes gene. 448(7153):591–594, 2007.

[41] W. Hoeffding. Scale-invariant correlation theory. *Masstabinvariante Korrelationstheorie, Schriften des Matematischen Instituts und des Instituts fÃijr Angewandte Mathematik der UniversitÃd't*, 5:179–233, 1940.

[42] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W.

Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167+, 2008.

[43] D. J. Hunter et al. A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874, May 2007.

[44] International HapMap Consortium, K. A. Frazer, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, Oct 2007.

[45] S. Jha, L. Kruger, and V. Shmatikov. Towards practical privacy for genomic computation. In *2008 IEEE Symposium on Security and Privacy*, 2008.

[46] J. Kennedy, I. MĂČndoiu, and B. PaÅ§aniuc. Genotype error detection using hidden markov models of haplotype diversity. *J Comput Biol*, 15(9):1155–1171, Nov 2008.

[47] Y. Kim, S. Feng, and Z. B. Zeng. Measuring and partitioning the high-order linkage disequilibrium by multiple order markov chains. *Genet Epidemiol*, 32(4):301–312, May 2008.

[48] J. E. Lunshof, R. Chadwick, D. B. Vorhaus, and G. M. Church. From genetic privacy to open consent. *Nat Rev Genet*, 9(5):406–411, May 2008.

[49] B. Malin. Re-identification of familial database records. In *AMIA Annu Symp Proc. 2006*, 2006.

[50] M. McCarthy, D. Smedley, and W. Hide. New methods for finding disease-susceptibility genes: impact and potential. *Genome Biology*, 4(10):119, 2003.

[51] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, May 2008.

[52] A. P. Morris, J. C. Whittaker, and D. J. Balding. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet*, 74(5):945–953, May 2004.

[53] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, New York, NY, USA, 2007. ACM.

[54] D. G. I. of Broad Institute of Harvard, L. U. MIT, N. I. of BioMedical Research, et al. Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science*, 316(5829):1331–1336, 2007.

[55] R. L. Plackett. Karl pearson and the chi-squared test. *International Statistical Review / Revue Internationale de Statistique*, 51(No.1):59–72, 1983.

[56] R. Robbins. Some applications of mathematics to breeding problems iii. *Genetics*, 3(4):375–389, 1918.

[57] N. E. Schulte. Statistical disclosure control (sdc) in practice: some examples in. *Statistical Journal of the United Nations Economic Commission for Europe*, 18(4):321–328, 2001.

[58] L. Scott et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, April 2007.

[59] R. Sladek et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, February 2007.

[60] J. Sun et al. Evidence for two independent prostate cancer riskâĂŞassociated loci in the hnf1b gene at 17q12. 40(10):1153–1155, 2008.

[61] H. C. Yang, C. C. Pan, C. Y. Lin, and C. S. Fann. Pda: Pooled dna analyzer. *BMC Bioinformatics*, 7:233–233, 2006.

[62] M. Yeager et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5):645–649, April 2007.

[63] I. H. D. Yuguo Chen and S. Sullivant. Sequential importance sampling for multiway tables. *The Annals of Statistics*, 34(1):523–545, 2006.

[64] E. Zeggini et al. Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science*, 316(5829):1336–1341, 2007.

[65] E. A. Zerhouni and E. G. Nabel. Protecting aggregate genomic data. *Science*, 322(5898):44a+, October 2008.

Table 6: Percentage of statistical power (at 0.05 type 1 error) left at various degree of missing values of LD measure r^C. r with absolution values smaller than t are missing. The simulation was done using the same parameters as used in generating Table 5.

Threshold t	0.5	0.2	0.1	0.01	0.001
$\%$ r values left	0.29	1.4	14	76	83
$\%$ power π left	8.5	24	79	100	100

APPENDIX

1. Proof of the validity of the sum statistics Given the definition of the signed allele correlation r,

$$r^C_{ij} = \frac{C_{11}C_{00} - C_{01}C_{10}}{\sqrt{C_{1*}C_{0*}C_{*1}C_{*0}}} \tag{6}$$

we have

$$\frac{\partial T^r_{ij}}{\partial C_{00}} = \frac{\partial r^C}{\partial C_{00}} \cdot Y^{00} = \frac{C_{11}\cdot[1-\frac{1}{2}\cdot(1-\frac{C_{10}C_{01}}{C_{11}C_{00}})\cdot(\frac{C_{00}}{C_{0*}}+\frac{C_{00}}{C_{*0}})]}{\sqrt{C_{1*}C_{0*}C_{*1}C_{*0}}} \geq 0 \tag{7}$$

for any pair of SNPs i and j. Similarly, we can get $\frac{\partial T^r_{ij}}{\partial C_{11}}$, $\frac{\partial T^r_{ij}}{\partial C_{10}}$ and $\frac{\partial T^r_{ij}}{\partial C_{01}}$ are also non-negative for any pair of SNPs i and j.

2. Proof of the optimality of the T_r statistics

Given $r = \{\hat{r}_{ij}|i,j\}$ and assuming the signed allele correlations follow the normal distribution, the optimal statistic for H_0: victim is in Case vs. H_A: victim is not in Case is $T_r = \Sigma_{i,j}\{2\mu_0(hatr^C_{ij} - r^C_{ij}) + \frac{1}{n}\{[\hat{r}^C_{ij} - (\mu_0 + r^C_{ij})]^2 - [\mu_0^2 + \frac{(1-(r^C_{ij})^2)^2}{n-1}]\}\}$. Applying the Neyman-Pearson lemma, where $\mu_0 \approx \frac{\partial r^C_{ij}}{\partial C_{00}}Y_{00} + \frac{\partial r^C_{ij}}{\partial C_{01}}Y_{01} + \frac{\partial r^C_{ij}}{\partial C_{10}}Y_{10} + \frac{\partial r^C_{ij}}{\partial C_{11}}Y_{11}$, n is the number of individual in the case group, r^C_{ij} is the theoretical signed allele correlations, and $\hat{r}^C_{ij}$ is the estimated signed allele correlations measure. By ommiting the second terms in the formula, which is $1/n$ times smaller than the first term, and replacing $\frac{\partial r^C_{ij}}{\partial C_{**}}$ by their signs (the real values needs pairwise frequencies to compute, which are unknown here), we have $T_r = 2\Sigma_{i,j}[(Y_{00} + Y_{11} - Y_{01} - Y_{10}) \cdot (\hat{r}^C_{ij} - r^C_{ij})]$. In practice, we estimate $\hat{r}^C_{ij}$) by using a reference group, i.e. $\hat{r}^R_{ij}$).