To Release Or Not To Release: Evaluating Information Leaks in Aggregate Human-Genome Data

Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, XiaoFeng Wang

Indiana University, Bloomington

Abstract. The rapid progress of human genome studies leads to a strong demand of aggregate human DNA data (e.g, allele frequencies, test statistics, etc.), whose public dissemination, however, has been impeded by privacy concerns. Prior research shows that it is possible to identify the presence of some participants in a study from such data, and in some cases, even fully recover their DNA sequences. A critical issue, therefore, becomes how to evaluate such a risk on individual data-sets and determine when they are safe to release. In this paper, we report our research that makes the first attempt to address this issue. We first identified the space of the aggregate-data-release problem, through examining common types of aggregate data and the typical threats they are facing. Then, we performed an in-depth study on different scenarios of attacks on different types of data, which sheds light on several fundamental questions in this problem domain. Particularly, we found that attacks on aggregate data are difficult in general, as the adversary often does not have enough information and needs to solve NP-complete or NP-hard problems. On the other hand, we acknowledge that the attacks can succeed under some circumstances, particularly, when the solution space of the problem is small. Based upon such an understanding, we propose a risk-scale system and a methodology to determine when to release an aggregate data-set and when not to. We also used real human-genome data to verify our findings.

1 Introduction

With rapid advancement in genome sequencing technologies, human genomic data has been extensively collected and disseminated to facilitate human genome studies (HGS). A prominent example is genome-wide association study (GWAS) [4], a research technique that has been demonstrated to be highly valuable for identifying the genetic factors underlying common diseases. In a GWAS study, a group of participants with a disease/phenotype of interest (cases) are genotyped to compare the statistical features of their single-nucleotide polymorphisms (SNPs)[1] to those of the individuals without the disease/phenotype (controls). It is highly desired that the DNA data gathered during this process can be conveniently accessed by other researchers, which will greatly benefit the HGS community. Such data dissemination, however, needs to be balanced with the protection of participants' privacy, which is of paramount importance to this kind of research: for example, revealing the identity of a case individual in a GWAS relates her to the disease under the study, which can have serious consequences such as denial of access to health/life insurance, education, and employment. Prior research shows that raw DNA data (genotypes) is often too risky to publish even after removal of

[1] Common terminologies of genomics are summarized in Appendix A.

explicit identifiers (such as name, social security number, etc.), as de-anonymization of a participant's identity can happen through examining the genetic markers related to her observable features (a.k.a. phenotypes) [8][2]. What has been thought to be of low risk is *aggregate genome data*, such as *allele frequencies*, i.e., the frequencies of different SNP values, because such data covers an individual's information with that of others. As an example, the NHGRI/NIH used to make allele frequencies publicly available.

Aggregate data releases. A recent development in inference technologies, however, has completely changed the risk perception associated with the aggregate data. Particularly, Homer et al [42] discovered that the presence of an individual in a case group can be reliably determined from allele frequencies using the victim's DNA profile, which can be acquired, for example, from a single hair or a drop of blood. In response to this finding, the NIH swiftly removed all aggregate genome data from the public domain to protect the participants of HGS and avoid legal troubles [2]. Today, those who want to access the data have to file an application and sign an agreement, a complicated procedure that is time consuming. This becomes a hurdle to the dissemination of the data critical to HGS, and as a result, provokes intensive debates [10]: some researchers pointed out that the NIH may have overreacted, as the attack power achievable over at least some data-sets can be very limited [21, 56]. On the other hand, such agreement-based protection has been found to be insufficient, as confidential user information can still be derived from other public sources: a recent study [57] shows that even the test statistics (e.g., p-values, r-squares) calculated from allele frequencies and published in HGS papers give away a significant amount of information, in some cases enough for identifying participants or even recovering portions of their DNA sequences. To make things worse, HGS researchers typically receive little guidance on what they are not supposed to share. Oftentimes, fine-grained allele frequencies/test statistics can be directly acquired from the authors of HGS papers.

Our work. The current way aggregate human DNA data is handled indicates a disturbing lack of understanding of its privacy implication: such data have been both over-protected, which unnecessarily restricts their availability to the HGS researchers, and underprotected, which exposes the HGS participants to privacy threats. Crucial to the progress of the human genome research, therefore, becomes an in-depth study on how to evaluate the information leaks in the aggregate data and determine when they are safe to release, which also poses a challenge to the privacy researchers. This paper reports our research that makes a first step toward this end. We consider two types of common aggregate data, the allele frequencies for both individual SNPs and SNP pairs, and the test statistics derived from the frequencies. Such data is studied under two typical threats, *identification attack* that uses an individual's DNA profile to determine her relation with an aggregate data-set [42, 49, 57], and *recovery attack* that re-constructs individuals' SNP sequences from such data. We investigated the feasibility of these attacks on different data and the difficulty in executing the attacks. For example, a recovery attack is doomed to fail if its target data cannot be uniquely mapped to a set of SNP sequences,

[2] The NIH's guideline for sharing GWAS data [8] explicitly states "the NIH takes the position that technologies available within the public domain today, and technological advances expected over the next few years, make the identification of specific individuals from raw genotype-phenotype data feasible and increasingly straightforward".

and the effects of an identification attack are constrained by the statistical powers that can be achieved over a data-set. When these attacks were likely to succeed, we further analyzed their computational complexities, which are often intractable. However, we found that solving such problems in practice can often be facilitated by the unique background knowledge of human genomes, particularly, the availability of a *reference* population that resembles case individuals.

Based upon such an analytical study, we further explored the potential to build a *risk scale system* to guide the dissemination of the aggregate data. Under this system, data-sets are classified according to their susceptibility to a specific type of attack: those in the *Green zone* are deemed safe to release; those in the *Red zone* are considered to be too risky to publish; the *Yellow zone* contains the data whose privacy implications are yet to be determined. For example, an aggregate data-set can be placed in the *Green zone* if it corresponds to multiple sets of DNA sequences and the intersection among these sets is small. To justify that a data-set belongs to the *Red zone*, we need to show that it can be exploited by a known attack. For this purpose, we also present a novel technique that improves on an existing identification attack [57]. Our new attack does not rely on the *integer constraint* used in the prior approach and works even on the data-sets that contain significant noise.

Contributions. We summarize the contributions of this paper as follows:

•*Fundamental studies on information leaks in aggregate data.* We performed both information-theoretic and complexity analyses on the common threats to different types of aggregate data. Our research sheds light on the fundamental questions on whether an attack on a specific data-set is feasible and how difficult it can be. Of particular importance here is our consideration of the special features of human genomes, which, as we show in the paper, can have significant impacts on the answers to these questions.

•*Preliminary research on the risk-scale system.* We propose a risk-scale system to classify aggregate data and guide the release of such data. Our research, though preliminary, is the first attempt to evaluate the risk of information leaks in a broad spectrum of aggregate data, including both single and pair-wise allele frequencies and different test statistics. We also present a new attack that identifies an individual from the test statistics published by an HGS, which serves the purpose of identifying the datasets that should not be released.

2 Backgrounds and Assumptions

2.1 Aggregate Human-Genome Data

Our research has been conducted on two types of aggregate genomic data, *allele frequencies* and *test statistics*. Both are among the most valuable data to human genome research and are also most widely disseminated: for example, the former has been published by the NIH [7] and the latter are elaborated in every GWAS paper [25,47,50,58].

Each SNP has two alleles, encoded as 0 (major) or 1 (minor). Using this encoding scheme, the DNA profiles (containing the nucleotide sequences of the participants) of N individuals L SNPs, could be simply represented as a $N \times L$ matrix. Figure 1 gives an extremely small sample of encoded SNP profiles of 5 participants and 8 SNPs. The *single-allele frequencies* f_i^p of a SNP site are the frequencies of the site's 'alleles, and

the *pair-wise allele frequencies* f_{ij}^{pq} of a SNP pair represent the frequencies of site i and j of the four allele combinations: $pq \in \{00, 01, 10, 11\}$. Note that allele frequencies can be simply calculated from allele counts by dividing N (e.g. $f_{ij}^{pq} = C_{ij}^{pq}/N$).

From the allele frequencies, test statistics are often computed in different human-genome studies. Particularly, GWAS researchers utilize *association tests* to detect the SNPs related to the disease under the study. These tests compare the single-allele frequencies of the case population with those of the control population, in the hope of identifying the genetic marker of the disease. The significance of each SNP (i.e., the strength of its tie to the disease) is measured by a p-value. Typically, those with p-values below 10^{-7} are selected as putative markers. Such marker-disease associations can also be quantified using other test statistics such as odds ratios.

In addition to analyzing individual SNPs, a GWAS also examines the putative marker's associations with other SNPs in the same genetic locus, called *linkage disequilibrium* (LD) [48], which could also have a connection with the disease. LD of a locus is typically measured by the test statistics such as D' and r-square, which are calculated from pairwise allele frequencies of the locus. Sometimes, researchers further analyzed the allele combinations involving multiple correlated SNPs, i.e., *haplotypes*, which are inferred from *genotypes* through a class of *phasing algorithms* [1, 54, 55].

Figure 2 shows how to calculate these test statistics and some sample values for Figure 1, which are routinely published in HGS papers [29, 50, 52, 58]. Oftentimes, these papers include the p-values of hundreds of SNPs and figures that illustrate their LDs. More detailed information can also be acquired from the authors. In our research, we focused on p-values and r-squares, the two most-commonly reported test statistics.

2.2 Threats

The threats studied in our research include *identification attack* and *recovery attack*, two major privacy concerns in human genome research. The first identification attack on aggregate data has been proposed by Homer, et al [42], which requires availability of a SNP profile from the victim. The objective here is to determine the presence of an individual in the case group, so as to relate her to a disease. To this end, the attacker runs a statistic test that evaluates whether the victim's SNP profile is independent from the single-allele frequencies of the case population. Let $Y_j \in \{0, 1\}$ be the allele of SNP j in the profile, and $\hat{f}_j^0$ and f_i^0 be the major allele frequencies of that SNP in the case population and a reference population, respectively. Homer's attack measures the

	L
N	0 0 0 0 0 1 0 0
	0 1 1 0 1 0 0 0
	1 0 0 1 0 0 0 0
	1 0 0 0 0 0 0 1
	0 1 0 1 1 1 1 1

Fig. 1. A 0-1 encoded SNP profiles of $N = 5$ individuals and $L = 8$ SNPs

Data	Name	Sample Values or Formula
C_i	single allele count for SNP i (major)	$C_1 = 3, C_3 = 4$
C_{ij}^{pq}	pair wise allele counts for SNP i and j	$C_{12}^{10} = 2, C_{13}^{00} = 2$
C_{ij}^{p*}	single allele count for SNP i	$C_{12}^{1*} = 2$
r_{ij}^2	r-square, measures association and LD	$\frac{(C_{ij}^{00}C_{ij}^{11} - C_{ij}^{01}C_{ij}^{10})^2}{C_{ij}^{0*}C_{ij}^{1*}C_{ij}^{*0}C_{ij}^{*1}}$

Fig. 2. Routinely published data (single allele counts without superscript means major counts, e.g. $C_i = C_i^0$).

following distance:

$$D(Y_j) = |Y_j - f_j^0| - |Y_j - \hat{f}_j^0| \tag{1}$$

Under the assumption that the distributions of individual allele frequencies are identical in the case and the reference, the sum of $D(Y_j)$ across a large number of SNPs follows a normal distribution with a zero mean if the victim is not present in the case group. Otherwise, the sum becomes positive and significantly deviates from the mean. In their paper, the authors report identification of a case individual with a extremely low false positive rate, given 25,000 SNPs of the victim. This line of research has been followed by multiple research groups [21, 43, 49, 56, 57]. Particularly, Sankararaman, et al [49] utilized the likelihood ratio test to estimate the upper-bound of the identification power achievable on single-allele frequencies. They also built a tool called SecureGenome [11] to evaluate such a threat on different data sets.

Besides single-allele frequencies, pair-wise allele frequencies and test statistics were also found to leak out a substantial amount of information. In prior research [57] , it was found that the identification attack can happen to even the test statistics published in GWAS papers, through a statistical test based upon *signed* r values. Given N sequences of L neighboring SNPs in the genome, the signed r_{ij} between two SNPs i and j ($1 \leq i < j \leq L$) is defined 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 the sequences with allele p ($p \in \{0, 1\}$) at SNP i and allele q ($q \in \{0, 1\}$) at SNP j, and C^{p*} and C^{*q} are single allele counts. r_{ij} can be computed from r_{ij}^2 (Figure 2) except its sign. Like Homer's approach, the attack needs a reference population whose r values are denoted by r^R, in addition to the case population (r^C), and a SNP profile from the victim in which $Y_{ij}^{pq} \in \{0, 1\}$ indicates whether her SNP pair ij has a pair-wise allele pq. A test statistic T_r is thus constructed as follows:

$$T_r = \sum_{1 \leq i < j \leq N} (r_{ij}^C - r_{ij}^R) \cdot (Y_{ij}^{00} + Y_{ij}^{11} - Y_{ij}^{01} - Y_{ij}^{10}) \tag{2}$$

T_r is much more powerful than the statistical attacks on single allele frequencies [57], as it makes use of the relations among SNPs, the linkage disequilibrium, which contain much more information than individual SNPs. A problem here, however, is the need to know the signs, which is not typically released. They are determined in the prior work [57] by taking advantage of *integer constraints*, base upon the assumption that the published r-squares are calculated from allele counts (integers) and are not perturbed by noise.

The recovery attack aims at re-constructing the SNP sequences (i.e., haplotypes) used in an HGS: prior research [57] report a successful restoration of 100 sequences involving 174 SNPs on a locus from their single and pair-wise allele frequencies. Note that these frequencies can be estimated through reverse engineering the test statistics published in GWAS papers [57]. Compared with the identification attack, such an attack can be more difficult to succeed and consume much more computing resources. However, it does not rely on the DNA profile from the victim.

An ideal privacy goal here is *differential privacy* [30], which ensures that two aggregated datasets differing from each other by one individual's data have indistinguishable statistical features. An example when this happens is that the data from a very large

number of participants is aggregated so that the contribution of an individual becomes negligible. This privacy goal, once achieved, can defeat inference attacks using all kinds of background knowledge. However, this condition is known to be hard to satisfy in a practical system. For genomic data, the knowledge about the victim's DNA profile and a good reference population is deemed as a strong assumption in the adversary's favor [21, 56]. Based on such an assumption, we report our research on the feasibility and complexity of these two types of attacks on the two types of datasets, and the methodology to determine whether a specific set of data is safe to release.

2.3 Adversary Model

We consider a probabilistic polynomial time adversary who can not accomplish the task that needs exponential computing power, for instance, sampling an exponential space to determine a probability distribution over this space. Other than that, we assume the adversary has sufficient resources and perfect information at her disposal for individual attacks. Specifically, for the identification attack, we consider that the adversary has access to the victim's DNA profile and a good reference population with an allele distribution identical to that of the case population. This is the best resource such an attack can expect [42, 57]. For the attack involving test statistics, we assume that high-precision data is available, which affects the outcome of such an attack, as indicated in the prior research [57].

3 Case 1: Identification Threat to Allele Frequencies

For single allele frequencies, the statistical identification threat they are facing has been well studied [42]. More specifically, SecureGenome [49] is proposed to evaluate the identification risk of such data. For pairwise allele frequencies, one can utilize a near-optimal statistic proposed in [57] (T_r) to assess the identification power achievable over the dataset. We also developed a likelihood ratio test Λ, which is also near-optimal. Due to the space limitation of the paper, we move the description of the test to Appendix B.

4 Case 2: Recovery Threats to Allele Frequencies

Given a set of pairwise allele frequencies, a recovery attack aims at partially recovering the haplotype sequences of HGS participants, which is completely realistic according to prior research [57]. These sequences, once restored, can be used to re-identify these participants, a threat well recognized by the NIH (see Footnote 1 and [8]). This section reports a new methodology for determining the susceptibility of different allele-frequency data to such an attack.

4.1 The Problem

Figure 3 illustrates the recovery attack, in which the adversary attempts to recover a matrix, with each of its row vectors being a haplotype sequence, from the constraints of pairwise allele frequencies[3]. This problem can be formulated as a *haplotype matrix recovery problem* below:

[3] Note that the pairwise allele frequencies contain all the information of single allele frequencies.

Haplotype matrix recovery problem. Consider an $N \times L$ haplotype matrix M that represents N haplotype sequences over L SNP sites. The set of pairwise allele frequencies of M is denoted by $d = \{f_{ij}^{pq}\}$, where p and q are the allele types at SNP sites i and j, respectively. Note that there are in total $\binom{L}{2}$ such pairs among L SNPs. Let S be the space of M (the matrix), and D be the space of d (the pairwise allele frequency). Given d and N, the adversary intends to recover the haplotype matrix, that is, to find an M' in S, which is equivalent to M ignoring the order of their row vectors.

It is conceivable that in some cases (some pairwise allele frequency d) the problem has unique *solution*: that is, there exists a unique matrix M, disregarding the ordering of its rows, that satisfies the constraints imposed by d, whereas in some other cases, the problem has no solution (i.e. the pairwise allele frequencies are not *satisfiable*), and in the remaining cases, the problem has multiple solutions. Figure 4 illustrates an example that multiple solutions exists for a given d. If there are multiple solutions and the intersection of all the solutions is small, when an attacker gets one solution, she has low confidence if any of the sequence in his solution is indeed in the original haplotype matrix.

Challenges in risk classification. To determine the risk scale of a given frequency set d, we first find out whether it has multiple solutions. If this is true and the overlap among these solutions is sufficiently small, we can comfortably put the set in the Green zone. Unfortunately, this decision turns out to be extremely difficult to make, because several problems on the haplotype matrix recovery are computationally hard. Specifically, we found that:

Theorem 1. *Determining if there is a haplotype matrix for a given pairwise allele frequency set is NP-complete.*

Corollary 1. *Determining the number of haplotype matrices for a given pairwise allele frequency set is NP-hard.*

Conjecture 1. Determining if a solution is unique for a given pairwise frequency set is Co-NP-complete.

Corollary 2. *Recovering one haplotype matrix for a given pairwise allele frequency set is NP-hard.*

Corollary 3. *Determining if there exists a solution for a given pairwise allele frequency set that does not contain a given row vector is NP-complete.*

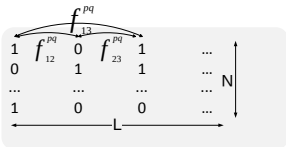

Fig. 3. Recovering a matrix from pairwise allele frequencies. Given a pairwise allele frequency set $d = \{f_{ij}^{pq}\}$, the attacker tries to recover the matrix satisfying d.

0 0 0	0 0 1
0 1 1	0 1 0
1 0 1	1 0 0
1 1 0	1 1 0
1 1 0	1 1 1

Fig. 4. The left matrix and the right matrix have exactly the same single allele frequencies and pairwise allele frequencies, but do not share any single haplotype sequence.

Corollary 4. *Recovering one haplotype matrix for a given pairwise allele frequency set that does not contain a given row vector is NP-hard.*

Proofs are provided in Appendix D. Theorem 1 to Corollary 4 show that determining the existence of unique or multiple solutions for a given allele frequency set and recovering even single one of them are all hard problems. Note that proving *average-case* complexity is well known to be difficult [37]. Nevertheless, our empirical study using IBM Cplex [5] with parallel enabled suggests that at least the decision problems here do not seem to be easy in the average time. We randomly sampled 10 matrices of size 100×80 and put them on a workstation with 4 Quad-Core Xeon 2.93GHz processors, none of them could be solved within one week.

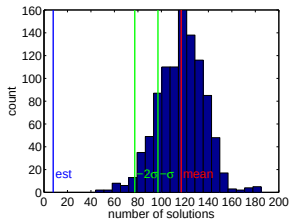

Fig. 5. Solution Distribution. (N = 40, L = 7, sample size = 1000, space ratio (estimated number of solutions) = 7.861, average = 116.855)

Fig. 6. Risk spectrum. When $\|S\| : \|D\| \gg 1$, data is placed in the Green zone. If there is a known attack, data must be placed in the Red zone. Otherwise further investigation is needed for the data (Yellow Zone).

Determination of risk scales. In spite of the difficulty in finding the number of solutions, it is still plausible to estimate whether a given frequency set is likely to have multiple haplotype matrix solutions, by considering solely the size of the recovery problem as determined by two parameters, i.e., the number of SNP sites L and the number of haplotype sequences N. We compare the solution space $\|S\|$ and the frequency set space $\|D\|$. When $\|D\| \approx \|S\|$, the corresponding frequency set is likely to have a unique haplotype matrix solution. Conversely, when $\|S\| \gg \|D\|$, a data-set d becomes very likely to have multiple solutions. Intuitively, the distribution of the solutions over the different d tends to have a very small deviation: that is, it is unlikely that only a few have many solutions while the others have unique ones. Furthermore, because the distribution is over a large number of variables (i.e. the elements in the haplotype matrix) and it is very complicated, the adversary cannot estimate the distribution without using exponential computing power. The adversary, who is unsure about the uniqueness of the solution, but, on the other hand, is aware of the strong indications that multiple solutions exist, will end up with little faith in any solution she is able to find. What is more, she may not even know how close to the real haplotype sequences her solution is, if $\|S\|$ becomes sufficiently large to ensure that many data-sets have multiple solutions.

Although it is difficult to rigorously define the distribution of solutions over d, we conducted an empirical study on the distribution as laid out below. We randomly sampled 1000 haplotype matrices of size $N = 40$ and $L = 7$, and calculated their pairwise

allele frequencies[4]. Using each set of these pairwise allele frequencies d as constraints, we computed for each instance all solutions that can be found by Cplex, a state-of-the-art NP solver [5] [5]. As expected, the distribution of the number of solutions is close to a normal distribution with a small standard deviation (Figure 5). The standard deviation (19) is on the same scale as the square root of the mean (116), indicating that it is unlikely that only a few d have many solutions while others have only a few or single solutions.

The above analysis indicates that we can have a shade-of-grey risk spectrum, as illustrated in Figure 6, which is approaching the Green end with the increase of the ratio $\|S\| : \|D\|$. Intuitively, this suggests that the larger the ratio, the less the adversary knows about the distance between her solution and the real one[6]. Upon the spectrum, we can use a distance threshold to determine when a frequency data-set can be designated to the Green zone. This research is elaborated in Section 4.2 and 4.3.

Towards the Red end of the spectrum, we proved that restoring a solution matrix from allele frequencies is NP-hard, even if the solution is known to be unique. However, we also acknowledge that the special features of human-genome data, particularly the LD relations among them, could make the problem tractable, as indicated in prior research [57]. Therefore, a conservative approach is to label a data-set "Red" only when it is found to be vulnerable to a known attack. Otherwise, the data-set is put in the Yellow zone, awaiting further investigation, if it is also not qualified for the Green zone. The details of this analysis are presented in Section 4.4.

4.2 When to Release

As discussed above, when the solution space becomes sufficiently larger than the space of allele-frequency sets, the threat of recovery attack can be mitigated, as the adversary cannot determine whether a given frequency data-set describes a unique set of SNP sequences. Here, we present an analysis on how large the solution space needs to be.

Solution-space analysis. Let us first consider the solution space S. For L SNPs, there are 2^L possible SNP sequences. The number of different solutions, each of which is an N by L haplotype matrix, is at least $\binom{2^L}{N}$, i.e., selecting N distinctive sequences from the 2^L sequences.

Then, we estimate the space of pairwise allele frequency sets D. Given N and a frequency set $d = \{f_{ij}^{pq}\}$, we can have a set of pairwise allele counts $\{C_{ij}^{pq}\}$, which directly determine the set of single allele counts $\{C_i\}$. Since for any SNP pair, the frequencies of one pairwise allele and one single allele are sufficient for inferring the frequencies of other alleles, pairwise or single, for the same SNP pair (see Inequality 3 in [57]), the set d is uniquely determined by $\{C_i\}$ and the set of pairwise major allele counts, which we denote by $\{C_{ij}\}$ for simplicity.

From the fact that C_{ij} and C_i can take any value in $[0, N]$ and there are $\binom{L}{2}$ SNP pairs and L single SNPs, we know that the number of different frequency sets d will not

[4] We chose this problem scale because L and N met the condition in Equation 3 and the problem is small enough to be solved by Cplex in reasonable time.

[5] We did not enumerate all putative solutions. Instead, we set the populate limit of Cplex as 200 to save memory and time. Hence, the number of solutions shown here may be smaller than the actual number of solutions.

[6] An exception here is some special cases, for example, when the frequencies of the pairwise allele type 00 become 1 for all SNP pairs. Such cases, however, can be identified before the data being released.

exceed $(N+1)^{\binom{L}{2}} \cdot (N+1)^L = (N+1)^{\binom{L}{2}+L}$. Comparing $\|S\|$ with $\|D\|$, we can get a necessary condition for the existence of multiple solutions: $\binom{2^L}{N} > (N+1)^{\binom{L}{2}+L}$. But it is too complex to use. Using Stirling's approximation, we get $\frac{2NL}{L(L-1)+2L}(1 - \frac{\log \frac{N}{e}}{L} - \frac{\log 2\pi N}{2NL}) > \log(N+1)$[7]. This gives us $\frac{2N}{L+1}(1 - \frac{\log \frac{N}{e}}{L} - \frac{\log 2\pi N}{2NL}) > \log(N+1)$. For $L > 200$, $1 - \frac{\log \frac{N}{e}}{L} - \frac{\log 2\pi N}{2NL} \approx 1$. Ignoring other constants, we get the following condition:

$$\frac{2N}{\log(N+1)} > L \tag{3}$$

For example, given $L = 200$, N should be greater than a threshold of 1000, which is roughly $5L$, to guarantee that Inquality 3 holds. Intuitively, this suggests that the adversary (with polynomial-time computing power) cannot uniquely determine the haplotype sequences in a case group of 1000 or more sequences (500 or more individuals) from the allele frequencies of 200 SNPs in an HGS study. Notably, even if N exceeds the threshold by a small margin, there will be a huge expansion of the ratio between $\|S\|$ and $\|D\|$. For example, increasing N by 1 (one more sequence in the case group) indicates the ratio of the spaces $\|S\|/\|D\|$ increases 2^L, which dramatically increases the likelihood of the presence of multiple solutions for the frequency set.

Partial recovery of haplotype matrix. The above analysis did not take into consideration the possibility that multiple solutions, although they exist, are close enough to each other for a given set of pairwise allele frequencies, e.g., there are a significant number of sequences shared between them. If this occurs and the attacker somehow recovered all the solutions (even though it is NP-hard, Corollary 1), and makes an intersection over these solutions, she knows the resulting common sequences must be in the case group. To defend against such attacks, we need stronger condition to assure the security of the pairwise allele frequency data to be released: for a specific haplotype sequence, there should exist another haplotype matrix solution that does not contain this sequence. When this happens, even if an attacker manages to obtain a solution (i.e. a set of haplotype sequences), she is not confident that *any* sequence in her solution is present in the actual haplotype matrix, because for any such sequence, there is always another haplotype matrix that is equally likely to be the actual matrix and also does not contain this sequence (although to find this matrix is NP-hard according to Corollary 4). Similarly, even if the attacker obtained multiple solutions, the intersection of these solutions will not give her any confidence that the sequence in the intersection must be present in the actual matrix.

To get this stronger condition, we consider the solution space for a given instance d with N rows (sequences) and L columns (SNP sites), but one haplotype sequence in the original matrix is not in these solutions. This is equivalent to the entire matrix space, i.e., $\frac{2^{N \times L}}{N!}$, subtracted by the matrix space with one fewer row (set as the given haplotype sequence), i.e., $\frac{2^{(N-1) \times L}}{(N-1)!}$. By using the same analysis from above, we get the following condition:

[7] Unless otherwise specified, $\log$ means $\log_2$ in this paper.

$$\frac{2(N-1)}{\log(N+1)} > L \tag{4}$$

Once the size of a haplotype matrix (N and L) meets this condition, its solution space will become sufficiently large that the intersection of all of its solutions is unlikely to contain even one haplotype sequence. This condition is also very close to that of Equation 1.

Empirical study. To verify whether the above privacy assurance is sufficient in practice, we conducted an empirical study on a number of small-scale problems. We randomly sampled 30 haplotype matrices that satisfy the condition (with $N = 40$ and $L = 8$), and for each haplotype sequence in the original matrix, we attempted to recover another haplotype matrix that did not contain this sequence but still has the same pairwise allele frequencies as those of the original matrix. Again, we used Cplex to search for all matrix solutions (with a populate limit of 200). In the end, for each of the haplotype sequences in the 30 matrices we sampled, at least 74 solutions were found that did not contain that sequence, indicating that given any haplotype sequence in a matrix, there likely exists an alternative solution (another haplotype matrix) associated with the pairwise frequency set of the original matrix, which does not include that sequence. This study shows that Equation 4 can be used to estimate when a pairwise frequency set is unlikely to be vulnerable to an intersection attack.

4.3 The Impact of Human Genetic Structure

A critical pitfall in the analysis above is that it does not take into consideration the prominent features of human genome sequences. Instead of being random binary sequences (0 for major and 1 for minor allele) as assumed in our model, human genome sequences contain complex structures that are well studied in human genetics and can be inferred from publicly accessible human genome data [6, 13]. Thus, the adversary could simply examine a solution she finds to determine whether it looks like a human genome sequence. This leads to the further reduction of the solution space $\|S\|$. In this section, we present another analysis based upon a human genetic model.

Human genetic model. We model haplotype sequences with a Markov chain (MC), a standard approach extensively used in human genetic research for the modeling of the LD structure (single and pairwise allele frequencies) in a specific genetic locus [38, 45, 46]. Given L SNP sites, the model can be represented as a heterogeneous Markov chain with a sequence of L *states* ($X_1X_2...X_L$), where $X_i \in 0, 1$, representing the major (0) or minor (1) allele, and an initial probability distribution (denoted by $P^0(X_1)$) as well as $L-1$ different transition probability matrices (denoted by $P^i(X_{i+1}|X_i)$) are used to model the transition probabilities from the i-th state to the $(i+1)$-th state, which are estimated from the single and pairwise allele frequencies using standard methods [38, 45, 46]. As a result, each of the 2^L haplotype sequences corresponds to a state sequence and the probability of observing it under the MC model can be computed by $P(X_1X_2...X_L) = P^0(X_1) \cdot \prod_{i=1}^{L-1} P^i(X_{i+1}|X_i)$. Once built from a group of haplotype sequences from human individuals (e.g. the case group or a reference group), the MC model can be used to evaluate the *effective* space of haplotype matrices that are likely sampled from real human individuals. Among totally 2^L possible haplotype sequences, the probabilities of observing some sequences are so low

that they are deemed unlikely to appear in human genomes, owing to the strong associations among neighboring SNPs. These sequences should not be considered when estimating the solution space of haplotype matrices. Assume the probabilities of 2^L sequences and a threshold θ (close to 1, e.g. 0.99999) are given, the effective space of haplotype sequences can be estimated by the number of most probable sequences that have a cumulative probability greater than θ. This was achieved in our research through an approximation algorithm presented in Appendix F.

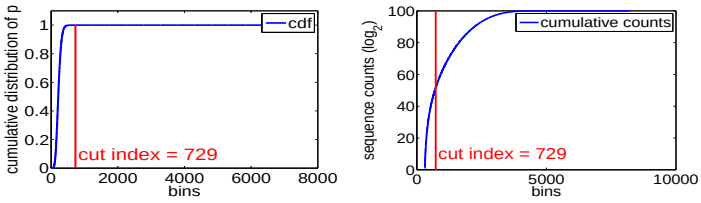

Fig. 7. The Markov Chain model for estimating the effective solution space. (A) Cumulative distributions of the probabilities of haplotype sequences, sorted in descending order of probabilities. Cutoff probability $\theta = 0.99999$. (B) Total number of most probable sequences vs. their cumulative probabilities. Vertical red line represent the cutoff.

Evaluation. To estimate the solution space under a human genetic model, we phased 3008 sequences from WTCCC ch7 of 100 SNPs by using PHASE [3]. We chose $2T = 8192$ bins to estimate distribution of haplotype sequences under the MC model. As shown In Figure 7 (A), with cutoff probability $\theta = 0.99999$, only 729 bins of $\approx 2^{52}$ most probable sequences are obtained, as compared to the entire space of 2^{100} haplotype sequences, which indicates that the incorporation of the human genetic model significantly reduces the effective space of haplotype sequences. Figure 8 shows the space comparison between $\|S\|$ and $\|D\|$. We could see that in the original analysis, we need about $5L$ sequences to ensure multiple solutions for the given pairwise allele frequencies. Defending against the intersection attack requires pretty much the same number of sequences as shown in the figure. To incorporate the human genetic model (the MC model), we need roughly $12L$ sequences.

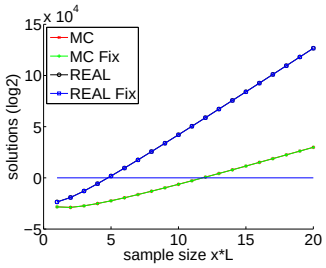

Fig. 8. Comparison between matrix space $\|S\|$ and constrain space $\|D\|$ for data fromWTCCC1 of SNP 100.

4.4 When Not to Release

For the frequency set that cannot be put in the Green zone, its solution is likely to be unique. The adversary who finds the solution has reason to believe that it is the correct one. Here, we elaborate how to classify such a data-set.

Red-zone data. Although recovering SNPs sequences is NP-hard in general, the special features of human genome can enable the attack to succeed on at least some frequency sets. Prior research reports a successful attack on a data-set related to 100 SNP sequences and 174 SNPs from the FGFR2 locus [57]. The approach leverages the LD relations among these SNPs to break the matrix into small blocks in a way that preserves the strong inter-SNP relations within individual blocks. Such relations allow the adversary to first restore individual blocks and then use the aggregated relations between blocks to connect them together.

To avoid releasing the data vulnerable to the recovery attack as well as overprotecting those that can actually be disclosed, we suggest to test a frequency set with known attacks and assign it to the Red zone when it is exploited. If the attacks fail, we can label the data-set as "Yellow" to leave the decision on its release to the future research.

5 Case 3: Identification Threat to Test Statistics

Besides allele frequencies, also widely disseminated by HGS are the test statistics computed from these frequencies. Particularly, HGS papers routinely report p-values and r-squares (r^2) over tens or even hundreds of SNP sites. Prior research [57] shows the key to an identification attack on such data is knowledge of the values of r or equivalently, their signs (given r^2). Once such information is given, we can use T_r [57] to decide whether a set of r-squares can be released, in the same way as SecureGenome and the Λ statistic (Section 3) do to single and pairwise allele frequencies. Specifically, we can release such a data-set if given all correct signs, the achievable statistical power on it, as reported by T_r, is still below a threshold. However, when the power turns out to be high, a decision to keep the data off limit can be premature: after all, there we assume that all the signs are recovered, which is by no means easy in practice, as discussed later in this section. Therefore, a question becomes how to seek a "tighter bound", allowing the statistics to be released when it is too difficult to recover a dangerous amount of information from them. This issue is addressed in this section.

The rest of the section presents our understanding of the problem: how sign recovery improves the chance of successful identifications and how difficult this can be done. Then, we come up with the yardsticks for releasing test statistics and describe a new potent attack that helps decide when data should be held from publication.

5.1 The Problem

How many signs need to be recovered? Prior research shows that under some circumstances, the signed r can be fully recovered [57]. However, there are plenty of real-world situations where this cannot be done. For example, when the statistics being released are of low precision, signs can only be partially restored [57]. Also, the existing attack only works under "integer constraints", when allele counts are used to compute the statistics, and relies on the consistent relations among SNPs: for example, the r-squares of

SNP 1 and 2, SNP 2 and 3, and SNP 1 and 3 must be consistent with each other. These conditions no longer hold in many studies that utilize a Maximum Likelihood Estimate (MLE) algorithm to roughly estimate r-squares from genotypes [53]. Therefore, it is completely realistic to consider the scenarios when the signs cannot be fully recovered.

An important question we are asking is how many correct signs a successful attack needs. The answer sheds light on the conditions under which the attack becomes ineffective. To find out the answer, we can analyze the relations between the rate of the correct signs used in an optimal test and the statistical power it can achieve on a particular data-set. Specifically, given a *rate* of correct signs α, we can randomly assign correct signs to the r of a fraction α of SNP pairs, and then run T_r under the assignment to determine its power, i.e., the rate of successful identifications. This test needs to be conducted repeatedly for each rate of correct signs, to get the maximum power under different sign assignments. In this way, we can obtain an estimated power-sign relation, and then use a threshold to determine the maximum rate of correct signs that will not pose a serious identification threat.

Complexity of releasing statistics. Given a threshold α ($\alpha \in [0, 1]$) of the correct sign rate, a set of test statistics (r-squares) can be placed in the Green zone if the adversary cannot correctly recover as many as α of all $\binom{L}{2}$ signs. This can be ensured if the set of r-squares is mapped to multiple sets of *valid* signed r values, and the overlap among these sets is below the threshold α. When this happens, the adversary, even if she can recover all these sets of signed r values, cannot identify enough signs with any confidence for a successful attack. Obviously, given $\binom{L}{2}$ r-squares over L SNP sites, there are totally $2^{\binom{L}{2}}$ possible *sign assignments*, with each of them corresponding to a different set of signed r values. However, not all of such assignments are valid: many of them do not correspond to any haplotype matrix, as those assignments lead to the r values inconsistent with each other.

We studied a *sign recovery problem*: given a set of r-square values r^2_{ij} over L SNP sites, a set of single allele frequencies f_i ($i = 1, 2, ..., L$), which could be recovered from p-values [57], and the total number of sequences in the case group (N), find a set of signed r values $\hat{r}_{ij}$ so that (1) $r^2_{ij} = \hat{r}^2_{ij}$; and (2) $\hat{r}_{ij}$ are *valid*, i.e. there exists a haplotype matrix whose pairwise allele counts C^{pq}_{ij} ($p, q \in 0, 1$) satisfy $N \cdot f_i = \sum_{q\in\{0,1\}} C^{0q}_{ij}$ for all i and j, and $r_{ij} = \frac{C^{00}_{ij}C^{11}_{ij} - C^{01}_{ij}C^{10}_{ij}}{C^0_i C^1_i C^0_j C^1_j}$. Similar to the *haplotype matrix recovery problem*, several key problems related to the sign recovery problem are computationally hard if we assume the haplotype matrix has more than just a few rows (haplotype sequences). This can be satisfied by all real HGS studies, which typically contains hundreds of individuals. Specifically, under this condition, we show that:

Theorem 2. *Determining if there exists a set of sign assignments of r for a given set of r-squares and single allele frequencies is NP-complete.*

Corollary 5. *Recovering a valid sign assignment for a given set of r-squares and single allele frequencies is NP-hard.*

Corollary 6. *Finding the number of valid sign assignment for a given set of r-squares and single allele frequencies is NP-hard.*

The proofs are provided in Appendix D. We note that these results have strong implications on classifying an r-square set into Green or Red zones. Briefly, an adversary

faces the following computational difficulty: assume that she manages to recover some sets of signs from r-squares, which itself is NP-hard; she still has no clue whether there are any other valid sign assignment and how many correct signs have been discovered in her solution. In other words, she will not have any reasonable confidence in the identification she makes from the r-square data-set. There is an exception, though: if the solution space of valid sign assignments (or equivalently their corresponding haplotype matrices) is sufficiently small, for example, as small as the space of r-squares, then the adversary has a good reason to believe that every set of r-squares has a unique valid sign assignment. Here the situation is analogous with that in Case 2 (Section 4). Similarly, we need a solution-space analysis to ensure that the adversary cannot get any useful information from a data-set to be released.

5.2 When to Release

Before placing a data-set to the Green zone, we need to ensure that the adversary cannot recover enough signs from it to achieve any significant identification power. Consider that a polynomial-time adversary learns from the ratio between the space of r-squares $\|R^2\|$ and the space of matrices $\|S\|$ that an r-square set can have κ solutions. Given a specific set of r-squares, she has no reason to believe that the set has fewer solutions, because she can neither determine the exact number of solutions nor sample the exponential space S (when N and L are large) to estimate the distribution of matrices over r-square sets. Also, recovering all these matrices is NP-hard and therefore the adversary has no clue how many different sets of valid signs exist, except that the number will not exceed κ. When κ is sufficiently large, the adversary, even after she manages to get a set of signs, does not know whether the overlap among all sets (which can be as many as κ) goes above $1-\alpha$ of all $\binom{L}{2}$ signs.

Solution-space analysis. Therefore, the condition for the release of an r-square set is that $\|S\| : \|R^2\|$ should be sufficiently large to ensure that the adversary does not know whether she recovers enough correct signs. As described in Section 4, $\|S\| \approx 2^{LN}(\frac{N}{e})^{-N}(2\pi N)^{-1/2}$. Since the space size of the r values is approximately $(N+1)^{\binom{L}{2}+L}$, from r to r-squares, the space shrinks to $\|R^2\| \approx \frac{(N+1)^{\binom{L}{2}+L}}{2^{\binom{L}{2}}}$. To ensure multiple solutions, we need $\|S\| > \|R^2\|$, which gives:

$$\frac{2N}{\log(N+1)-1} > L \tag{5}$$

For example, for a locus involving 100 SNPs, at least 225 individuals (450 haplotype sequences) should be in the case group to ensure the existence of multiple solutions. Not surprisingly, this is less stringent than the condition of placing a set of pairwise allele frequency in the Green zone (where one needs to have at least 500 sequences for a 100-SNP locus), because r-squares contain less information than the pairwise allele frequencies. To further prevent the adversary from identifying more than $1-\alpha$ of the correct signs, we need to make it possible to have an element in R^2 be mapped by at least $2^{(1-\alpha)\binom{L}{2}}$ elements[8] in S. To ensure this, we must have that $\|S\|$ is at least

[8] Note that the adversary has to consider the situation that all these elements (matrices) are associated with different r sets, as she has no computing power to estimate the relations between r and matrices.

$2^{(1-\alpha)\binom{L}{2}}$ times as large as $\|R^2\|$. This ultimately gives us the following condition:

$$\frac{2N}{\log(N+1)-1+\alpha} > L \tag{6}$$

Considering human models. Again, when the special properties of human genomes are being considered, we need to re-assess the matrix space $\|S\|$ based upon a human genetic model, as described in Section 4.3. In our research, we ran the approximation algorithm (Section 4.3) to identify L and N that satisfy the above conditions (multiple sets of signs with a large distance), using the WTCCC1 data.

Figure 9 shows the result of the experiment involving 100 SNPs. As we can observe from the figure, in absence of a human model, a population with more than 250 individuals (500 sequences) are required to make sure that no more than 60% of signs can be identified. If we consider the human features, we need a population of at least 600 individuals ($N > 1200$).

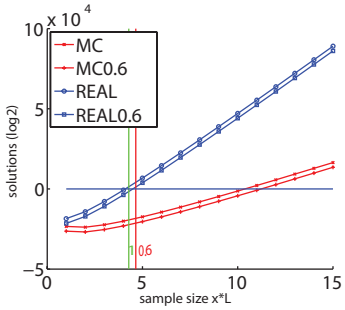

Fig. 9. Comparison between matrix space $\|S\|$ and $\|R^2\|$ for data from WTCCC1 of $L = 100$ SNP. Vertical line shows the required sample size estimated from formula 5 and 6 and then added by a buffer of $0.5L$.

5.3 When Not to Release

When the space of matrices S comes close to that of the r-squares, the adversary knows that once she acquires a set of valid r values, they are likely to be correct. Although we have shown that recovering signed r values from r-squares is NP-hard (Corollary 5), some instances of the sign recovery problem may be easy to solve, in particular when a human genetics model is employed to help solve the problem. Here we present a new attack technique that helps determine when this situation occurs, and thus a data-set should not be released.

A new attack. Although sign recovery from r^2 is NP-hard in general, the special properties of human-genome data could make the attack feasible under some circumstances, as demonstrated by prior research [57]. Therefore, we again need to do the Red-Yellow classification as described in Section 4.4: a set of r-squares is labeled as Red if it is found to be vulnerable to a known attack, and Yellow otherwise. This classification relies on the availability of the testing techniques, the most potent attacks we can find, for evaluating the amount of the information inferable from a data-set.

Prior research [57] reports such an attack technique, which utilizes the integer constraints and high precision statistics to recover r from r^2. In this paper, we present a new attack technique that works on the r^2 data directly estimated from genotypes with MLE, as did in many human-genome studies [53]. Such data are of poor quality, in the sense that they do not reflect the real allele counts and are not even consistent with each other. This makes sign recovery difficult.

The new techniques we propose leverage the special structure of human genome sequences and the availability of a reference population. Different groups of people often share very similar linkage disequilibrium structures. Thus, a haplotype sequence can be viewed as a combination of many small blocks of SNP sites (called haplotype blocks), with strong linkages among the SNPs within the same block and weak linkages among those in different blocks. Such relations are captured by aggregate LD information, r-squares in particular. Biologically, this property is caused by the fact that human DNA is passed from one generation to another in a way that recombinations occur more often at some regions in the genome than other regions. Those hot spots of recombinations form the boundaries of haplotype blocks. As a result, a reference population often has a very similar block structure as the case group. Our approach adjusts the individual haplotype blocks of the reference population as well as their combinations using a machine learning algorithm, in attempt to make the r-squares of the resulting haplotype matrices (of the reference) as close to that of the case group as possible. Due to the limited space of the paper, we put the details of the attack in Appendix C.

Evaluations. We ported the `LD` function, which is used in many GWAS papers for calculating MLE r^2, from the snp.plotter [12] package of `R` [9] to Matlab and implemented the recombination attack using a stochastic hill climbing algorithm with multiple starting points. Then, we evaluated the attack on the data extracted from WTCCC1. We extracted 180 SNPs from chromosome 7. A case group and a reference group of 100 each were randomly sampled from the data-set. After that, the MLE-estimated r^2, together with single allele frequencies, was used as the optimization target for both inner block and inter block recombinations. On average, the sign agreement rate between the initial haplotype matrix (reference) and the target matrix (case) was 58.7%, which had very small power (identification rate 3.0% under a false positive rate 1%). After learning, the sign rate agreement was improved to 67.2% on average and the identification rate became 8.1%: that is, our approach enabled an adversary to identify about 8 participants from the aggregate data with a poor quality.

6 Case 4: Recovery Threats to Test Statistics

Previous research has shown that single allele frequencies can be accurately recovered from these published statistics [57]. In this section, we study the possibility for an attacker to recover haplotype matrices from published r-square values and single allele frequencies. As discussed in Section 5, since r-squares contain less information than pairwise allele frequencies, such a recovery attack is more difficult than the attack discussed in Section 4, though it is not completely impossible.

When to release. Conservatively, we can first check whether r-squares can be released using the method discussed in Case 2, assuming that the attacker has a way to obtain pairwise allele frequencies from the r-squares. It is obvious that if the pairwise allele

frequencies are placed in the Green zone, the r-squares should be safe to release as well. A more strict bound can be found shown in 7. we left the analysis in Appendix E.

$$\frac{2(N-1)}{\log(N+1)-1} > L \tag{7}$$

When not to release. However, as shown in the prior research [57], it is plausible to fully recover r with proper algorithms under some circumstances. When this happens and Inequality 7 is not satisfied, the adversary could gain some confidence in the outcome she acquires. Therefore, like in other cases, we recommend to use the method described in [57] to verify if a data-set is indeed vulnerable and needs to be placed in the Red zone.

7 Related Work

The problem of releasing aggregate data while preserving their privacy has been extensively studied in privacy preserving data analysis [30, 35], statistical disclosure control [18, 19, 33], inference control [24] and privacy-preserving data mining [14, 15]. However, the properties of human genome data make the problem special in this domain, which has not been well investigated. Especially, human individuals share about 99.9% genomic sequences, which makes it easy to find a reference group from public sources such as HapMap [6]. This enables both Homer's attack and the statistical attack proposed in [57], as elaborated in Section 2.2. Also remotely related to our research is the work on privacy preserving genome computing [16, 22, 44], which however does not focus on protecting the outcomes of a computation from being inferred.

The recent progress in human genome research [32, 39] has made a great demand on convenient access to sensitive human genome data for research purpose. The problem of balancing privacy protection and data sharing in this domain, however, has not been seriously studied until Homer, et al. published their findings [42] a couple years ago. After that, several research groups, including us, have started working on this important issue [21, 43, 49, 56, 57]. As a prominent example, Sankararaman, et al [49] recently propose a technique (SecureGenome) for measuring the maximum statistical powers achievable on a set of single-allele frequencies. Most of these studies focus on single allele frequencies, which has been found in prior research to be insufficient [57], as sensitive information can also be inferred from other sources like test statistics. The research presented in this paper is the first attempt to understand and assess the risk in releasing different types of aggregate data, under typical inference threats.

Recovering SNP sequences is related to the research on contingency table release [20, 23, 28, 41, 59], and discrete tomography [40], which tries to reconstruct a matrix from a small number of projections. However, the specific problem of restoring a matrix from pair-wise allele counts is new, up to our knowledge, and the related complexity problems have not been studied before.

The Red-zone data identified by our techniques are not supposed to be released directly. However, they could still be published after proper sanitization and obfuscation. Such techniques have been studied in data-based privacy [17, 19, 31]. Particularly, the privacy policy based upon *Differential privacy* [30], once enforced, can make an identification impossible. Therefore, an important research direction is to develop effective techniques to achieve such a privacy objective on aggregate human genome data.

8 Conclusion

Availability of aggregate human DNA data is of great importance to human genome studies. Recent research shows that such data are vulnerable to different types of privacy threats, which could lead to identification of the participants of these studies and disclosure of their sensitive genetic markers. Therefore, a critical question becomes how to evaluate such a risk and determine when the data are safe to release. In our research, we make the first attempt to answer this question. We identified the problem space of aggregate data release, considering both different types of data available in the public domains (allele frequencies and test statistics) and common threats to such data (identification attack and recovery attack). Through a systematic exploration of the space, we gained an important new understanding of the problem. Specifically, we found that inferring useful information from such data is difficult in general: the adversary often does not have enough information and needs to solve NP-complete or NP-hard problems. On the other hand, we also show that an attack can still happen under some circumstances, particularly when the solution space of the problem is small. Based upon such an understanding, we propose a new risk-scale system that determines when data can be safely released, through analyzing their solution spaces. The important findings of this research are summarized in Table 1.

Table 1. Summary of Main Results

	Identification Attack	Recovery Attack
Aggregate data: single allele frequency pairwise allele frequency	1. Release if the power of Λ and T_r is very low. 2. NOT to release if the power of Λ and T_r is higher than a threshold.	1. Release if $\frac{2(N-1)}{\log(N+1)} > L$ adjusted by MC model. 2. NOT to release if a known attack can recover matrix.
Statistics: r^2, p-value	1. Release if $\frac{2N}{\log(N+1)-1+\alpha} > L$ adjusted by MC model. 2. NOT to release if T_r statistic succeeds.	1. Release if $\frac{2(N-1)}{\log(N+1)-1} > L$ adjusted by MC model. 2. NOT to release if a known attack succeeds.

Given the scale and the depth of this data-release problem, many open issues remain in the problem space. Particularly, in Case 1, our test for measuring the statistical power on pair-wise allele frequencies is not optimal. Actually, design of an optimal test here seems to be extremely challenging, as these frequencies are related, and their joint distribution is very complicated and very difficult to model. For other cases, an important issue is how to narrow the range of the Yellow zone, to get tighter bounds for releasing or not releasing an aggregate data-set. Also important is the study on new anonymization techniques that obfuscate the Red-zone data to achieve differential privacy without substantially compromising their scientific value.

References

1. Haplotype Estimation and Association. http://slack.ser.man.ac.uk/theory/association_hap.html, 2005.

2. NIH Background Fact Sheet on GWAS Policy Update. http://grants.nih.gov/grants/gwas/background_fact_sheet_20080828.pdf, 2008.
3. fastPHASE. http://stephenslab.uchicago.edu/software.html, 2010.
4. Genome-Wide Association Studies. http://grants.nih.gov/grants/gwas/, 2010.
5. Ibm ilog cplex optimizer. http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/, 2010.
6. International HapMap Project. http://www.hapmap.org/, 2010.
7. National Institutes of Health. http://www.nih.gov/, 2010.
8. 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, 2010.
9. The R project for statistical computing. http://www.r-project.org/, 2010.
10. Re-identification and its discontents. http://www.genomicslawreport.com/index.php/2009/10/13/re-identification-and-its-discontents/, 2010.
11. SecureGenome. http://securegenome.icsi.berkeley.edu/securegenome/, 2010.
12. SNP.plotter. http://cbdb.nimh.nih.gov/~kristin/snp.plotter.html, 2010.
13. Wellcome Trust Case Control Consortium (WTCCC1). https://www.wtccc.org.uk/ccc1/, 2010.
14. 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.
15. R. Agrawal and R. Srikant. Privacy-preserving data mining. *SIGMOD Rec.*, 29(2):439–450, 2000.
16. 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.
17. 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.
18. L. L. Beck. A security machanism for statistical database. *ACM Trans. Database Syst.*, 5(3):316–3338, 1980.
19. 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.
20. C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8(1):3–62, 1936.
21. R. Braun, W. Rowe, C. Schaefer, J. Zhang, and K. Buetow. Needles in the haystack: Identifying individuals present in pooled genomic data. *PLoS Genet*, 5(10):e1000668, 10 2009.
22. 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.
23. 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.
24. F. Y. Chin and G. Ozsoyoglu. Auditing and inference control in statistical databases. *IEEE Trans. Softw. Eng.*, 8(6):574–582, 1982.

25. A. Chiò, J. C. Schymick, et al. A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. *Hum Mol Genet*, 18(8):1524–1532, Apr 2009.
26. V. Chvatal. Recognizing intersection patterns. In *Combinatorics 79, Part I*, pages 249–251. North-Holland Publishing Company, 1980.
27. D. P. Dailey. Uniqueness of colorability and colorability of planar 4-regular graphs are np-complete. *Discrete Mathematics*, 30(3):289 – 293, 1980.
28. 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.
29. R. H. H. Duerr et al. A genome-wide association study identifies il23r as an inflammatory bowel disease gene. *Science*, October 2006.
30. C. Dwork. Differential privacy. In *in ICALP*, pages 1–12. Springer, 2006.
31. C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, 2006.
32. A. O. Edwards, R. Ritter, et al. Complement factor H polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, Apr 2005.
33. 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.
34. M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.
35. 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.
36. O. Goldreich. Computational complexity: A conceptual perspective, 2006.
37. O. Goldreich and S. Vadhan. Special issue on worst-case versus average-case complexity editors' foreword. *Comput. Complex.*, 16:325–330, December 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 et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, Apr 2005.
40. G. T. Herman and A. Kuba. *Advances in Discrete Tomography and Its Applications (Applied and Numerical Harmonic Analysis)*. Birkhauser, 2007.
41. W. Hoeffding. Scale-invariant correlation theory. *Masstabinvariante Korrelationstheorie, Schriften des Matematischen Instituts und des Instituts fr Angewandte Mathematik der University*, 5:179–233, 1940.
42. N. Homer et al. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167+, 2008.
43. K. B. Jacobs et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41(11):1253–1257, October 2009.
44. S. Jha, L. Kruger, and V. Shmatikov. Towards practical privacy for genomic computation. In *2008 IEEE Symposium on Security and Privacy*, 2008.
45. 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.
46. 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.
47. F. Renström et al. Replication and extension of genome-wide association study results for obesity in 4,923 adults from northern sweden. *Hum Mol Genet*, Jan 2009.

48. R. Robbins. Some applications of mathematics to breeding problems iii. *Genetics*, 3(4):375–389, 1918.
49. S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin. Genomic privacy and limits of individual detection in a pool. *Nat Genet*, 41(9):965–7, 2009.
50. L. Scott et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, April 2007.
51. J. Simon. On the difference between one and many (preliminary version). In *Proceedings of the Fourth Colloquium on Automata, Languages and Programming*, pages 480–491, London, UK, 1977. Springer-Verlag.
52. R. Sladek et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, February 2007.
53. M. Slatkin and L. Excoffier. Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity*, 76 (Pt 4):377–383, Apr 1996.
54. M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American journal of human genetics*, 73(5):1162–1169, November 2003.
55. M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4):978–989, April 2001.
56. P. M. Visscher and W. G. Hill. The limits of individual identification from sample allele frequencies: Theory and statistical analysis. *PLoS Genet*, 5(10):e1000628, 10 2009.
57. R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. In *CCS '09: Proceedings of the 15th ACM conference on Computer and communications security*, pages 534–544, New York, NY, USA, 2009. ACM.
58. 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.
59. I. H. D. Yuguo Chen and S. Sullivant. Sequential importance sampling for multiway tables. *The Annals of Statistics*, 34(1):523–545, 2006.

A Terminologies

B Likelihood Ratio Test

Assume the pairwise allele frequencies over L SNP sites are given for a case group C and a reference group R, denoted as f_{ij}^{pqC} and f_{ij}^{pqR}, respectively, where $1 \leq i < j \leq L$ represent two different SNP sites and $p, q \in 0, 1$ represent the allele type at the sites i and j, respectively. To test if an individual with the haplotype sequence s_k ($1 \leq k \leq L$) is likely to be within the case group [9], we design the following likelihood ratio test under the null hypothesis (H_0) that s_k is sampled from pairwise allele distribution defined by f_{ij}^{pqC}, and the alternative hypothesis (H_1) that s_k is sampled from pairwise allele distribution defined by f_{ij}^{pqR}.

$$\Lambda(s) = \frac{P(s|f^C)}{P(s|f^R)} = \prod_{ij}[\frac{f_{ij}^{s_i s_j C}}{f_{ij}^{s_i s_j R}}] = \prod_{ij}[\prod_{pq}([(\frac{f_{ij}^{pqC}}{f_{ij}^{pqR}})^{Y_{ij}^{pq}}])] \tag{8}$$

$$\log(\Lambda) = \sum_{ij}\sum_{pq}[Y_{ij}^{pq} \cdot (\log f_{ij}^{pqC} - \log f_{ij}^{pqR})] \tag{9}$$

[9] In practice, each human individual carries two haplotype sequences, which can be tested independently.

Table 2. HGS Terminologies used in this paper.

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

where Y_{ij}^{pq} is the indicator for the pairwise allele types of the individual haplotype at SNP sites i and j: $Y_{ij}^{pq} = 1$ if $p = s_i$ and $q = s_j$; $Y_{ij}^{pq} = 0$ otherwise. The statistic Λ provides another test to evaluate the identification risk associated with a set of pairwise allele frequencies, assuming a perfect reference group can be obtained, as SecureGenome does on single allele frequencies. We note that Λ is not necessarily the optimal statistical test on a given set of pairwise allele frequencies of case and reference group. Even though the likelihood ratio test is known to be nearly optimal for the case and reference pairwise allele frequencies of a specific SNP pair, the pairwise allele frequencies cannot be treated as independent random variables (instead they are highly correlated), and thus the summation over all pairwise allele frequencies do not guarantee the overall optimality of the statistic test. We conjecture that it is very hard to design an optimal statistic for distinguishing two sets of pairwise allele frequencies.

C Haplotype Recombination Attack

Let $H_i = [h_i^1, h_i^2, \cdots, h_i^{k_i}]_{1 \le i \le K}$ be the haplotype blocks of a locus and h_i^j be the haplotype sequences within block i, K be the total number of blocks, and k_i be the number of haplotype sequences within block i. Assume haplotype sequence h_i^j occurs at a frequency of f_i^j in human population. The attack consists of three steps. First, we determine the block structure, the haplotype sequences in each block (h_i^j) and their frequencies (f_i^j) from a reference group of haplotype sequences at the same locus. Afterwards, we initialize a haplotype matrix of the expecting size by randomly combined haplotype sequences in each block with the expected frequencies. Next, we refine the frequencies of haplotype sequences in each block. We use MLE-estimated

r^2 and single allele frequencies to devise a target function and use a stochastic hill climbing (SHC) algorithm for the optimization. Specifically, our goal is to minimize $E = \beta\|\hat{r}_i^2 - r_i^2\|/\binom{L_i}{2} + (1-\beta)\|\hat{c}_i - c_i\|/L_i$, where r_i^2 and c_i represent the r-squares and single allele frequencies in block i, and $\hat{r}_i^2$ and $\hat{c}_i$ represent the r-squares and single allele frequencies of the haplotype matrix being optimized. After this step, we get a refined frequencies of haplotype sequence in each block. Finally, we attempt to refine the combination of the haplotype sequences in different blocks to fit the r-squares of the case group between SNP sites in different blocks. We rearrange the order of the haplotype sequences in each block. Due to the inconsistency in the MLE estimated r-squares especially for the long distance blocks, we can not satisfy all r-squares. The rearrangement process can be viewed as another optimization procedure that tries to minimize the distance between the r-squares computed from the haplotype matrices and the target r^2 for the pairs of SNP sites in two different blocks. We applied the similar algorithm to the optimization as in the previous step and ultimately we get a final haplotype matrix.

D Proofs of Theorem 1, Corollary 1, Corollary 2 and Corollary 4 Conjecture 1, Theorem 2, Corollary 5, Corollary 6

Proof of Theorem 1 This problem can be formalized as an existence problem EP(C_i, C_{ij}) which is to determine whether an $N \times L$ binary matrix M exists that satisfies the constraints of the sets of single allele counts $\{C_{i\in[1,L]}\}$ (e.g. the number of 0s in column i) and pairwise major allele counts $\{C_{ij\in[1,L]}\}$ (e.g. the number of 00 pairs of column i and column j). NOTE that these two sets are equivalent to the set of pairwise allele frequencies and may be used interchangeably in this paper. e.g. $C_i = 3$, termed as 3-EP(C_i, C_{ij}). Consider a special case of EP(C_i, C_{ij}), denoted as 3-EP, where all given single allele counts are 3 ($C_i = 3$). We prove 3-EP is NP-complete by reducing the 3-*Recognizing Intersection Patterns Problem*(3-RIPP(A)), a known NP-complete problem [26] to it. 3-RIPP(A) is described as: given $A = [a_{ij}]_{L\times L}$ in which $a_{ii} = 3$, is there an integer set collection $H = \{H_1, H_2, \cdots, H_L\}$ such that $a_{ij} = |H_i \cap H_j|$ for $1 \le i, j \le L$. Obviously, 3-EP $\in$ NP. Given an arbitrary instance of 3-RIPP(A), we construct an instance of 3-EP(C_i, C_{ij}) by setting $C_{ij} = a_{ij}$ for $1 \le i \ne j \le L$ and setting $C_i = a_{ii}$ for $1 \le i \le L$. Suppose $M_{N\times L}$ is a solution of EP. We can convert each column of $M_{N\times L}$ into a set, where the row indices of 1s in the i-th column form the elements in the set H_i. Therefore, We get $|H_i \cap H_j| = a_{ij} = C_{ij}$ for $1 \le i, j \le L, i \ne j$ and $|H_i \cap H_i| = a_{ii} = C_i = 3$. So $\{H_i\}$ represent a solution of 3-RIPP(A). Conversely, suppose $H = \{H_1, H_2, \cdots, H_L\}$ is a solution of 3-RIPP(A). We can construct a solution M of 3-EP by converting each set H_i into a column of length L where for each element $k \in H_i$, fill in the k-th position by 1 in the i-th column of $M_{N\times L}$, and all the other positions by 0. Clearly the resulting matrix $M_{N\times L}$ is consistent with (C_i, C_{ij}), and thus is a solution of 3-EP. Because the conversions described above can be done in polynomial time, 3-EP(C_i, C_{ij}) is NP-complete. Therefore, EP(C_i, C_{ij}) is also NP-complete since its special case 3-EP(C_i, C_{ij}) is NP-complete.

Proofs of Corollary 1, Corollary 2 and Corollary 4 Given the number of haplotype matrices for a given pairwise allele frequency set d, we can directly determine if there is a haplotype matrix for d. So finding the number of solutions for given d is NP-hard.

By Theorem 1, EP is NP-complete. So the corresponding searching problem of recovering matrix from pairwise allele frequencies is NP-hard.

For Corollary 4, a matrix recover problem of size $N \times L$ can be extended by adding one column of all 1. Besides the pairwise allele frequency constraints, we add an additional constraint excluding a row vector v of length $L+1$ with the last bits as 0. Note that this constraints will never affect the solution of the new matrix since the last bits of any sequence of the new matrix can not be 0. If the new matrix has a solution, the original problem has a solution by removing the last column.

Argument of Conjecture 1. Given graph G and a k-coloring of G, to determine whether a given k-coloring of G is unique is Co-NP-complete [27]. According to a conjecture [36] that the solution counting problem for all known NP-complete problems listed in [34], including the graph k-colorability problem and 3-RIPP(A), are #P-complete. Thus, there should be a parsimonious reduction [51] from the graph k-colorability problem to the 3-RIPP(A), and the reduction we constructed in the proof of Theorem 1 from 3-RIPP(A) to 3-EP is also a parsimonious reduction. These two reductions ensure every instance of the graph k-colorability problem has the same number of solutions as the corresponding instance of 3-EP. So if we can determine whether a given solution is unique for an arbitrary instance of 3-EP by an algorithm, we can solve the same problem for the graph k-colorability problem by the same algorithm. Thus we can conclude that to determine whether a given solution is unique for an arbitrary instance of 3-EP is also Co-NP-complete, and the same conclusion also holds for the general EP problem.

Proofs of Theorem 2, Corollary 5. The problem stated in Theorem 2 is equivalent to the problem of determining if there is a matrix satisfying (r^2, C_i, N), denoted as SEP(r^2, C_i, N).

We first prove a special case case of SEP(r^2, C_i, N) where $C_i = 3$ for all $1 \leq i \leq N$ and $N > 18$, denoted as 3-SEP(r^2, C_i, N). we prove 3-SEP is NP-complete as below. Obviously, 3-SEP $\in$ NP. We prove 3-SEP is NP-complete by constructing a Karp reduction from 3-EP(C_i, C_{ij}). Given an arbitrary instance of 3-EP(C_i, C_{ij}), we construct an instance of the 3-SEP(r^2, C_i, N) as follows. Given C_i, C_{ij}, if $\sum_1^N C_i > 18$, we set $N = \sum_1^N C_i$, else set $N = 19$. Note that setting $N > \sum_1^N C_i$ will not affect the existence of a solution since we can always add arbitrary rows of all 0s to a solution of matrix without changing its single allele and pairwise allele counts. We then set $r_{ij}^2 = \frac{(C_{ij}N - C_iC_j)^2}{C_i(N-C_i)C_j(N-C_j)}$. This conversion can be done in polynomial time. We now prove that given a solution $M_{N\times L}$ of a 3-SEP problem, $M_{N\times L}$ is also a solution of the corresponding problem of 3-EP and vice versa. This is because, for $N > 18$ and $C_i = 3$, the function $r_{ij}^2 = \frac{(C_{ij}N - C_iC_j)^2}{C_i(N-C_i)C_j(N-C_j)}$ is bijective. It is easy to verify that the function is injective (from C_{ij} to r_{ij}^2). We now prove the inverse function is also injective. Assume there are two different C_{ij}, denoted as $C_{ij}^{'}$ and $C_{ij}^{''}$ give the same r_{ij}^2. We then must have $\frac{(C_{ij}^{'}N - C_iC_j)^2}{C_i(N-C_i)C_j(N-C_j)} = \frac{(C_{ij}^{''}N - C_iC_j)^2}{C_i(N-C_i)C_j(N-C_j)}$. Since $C_i = C_j = 3$ and $C_{ij}^{'} \neq C_{ij}^{''}$, then we get $(C_{ij}^{'} + C_{ij}^{''})N = 18$. However, because $C_{ij}^{'} + C_{ij}^{''} \geq 1$ since they are not equal and $N > 18$, we have $(C_{ij}^{\prime} + C_{ij}^{''})N > 18$, which is an contradiction. So there is a bijection between 3-SEP and 3-EP. Therefore, we conclude that the SEP

problem is NP-complete since it's special case 3-SEP is NP-complete. So the problem stated in Theorem 2 is also NP-complete. Also, the corresponding searching problem of recovering valid sign assignment from r^2 and single allele frequency is NP-hard.

Proof of Corollary 6. Given the number of valid sign assignments for given (r^2, C_i, N), we can directly determine if there is a valid sign assignment. So finding the number of valid sign assignment for given (r^2, C_i, N) is NP-hard.

E Case 4: more strict bound

To obtain a more specific guideline for releasing r-square values, we adopt a similar approach as described in Section 4, which compares the haplotype matrix space $\|S\|$ with the r-square space $\|R^2\|$. If $\|S\| > \|R^2\|$, it is very likely that multiple haplotype matrix solutions exist for a given set of r-squares. We have shown that to recover a haplotype matrix from r-squares is NP-hard, and to determine the number of haplotype matrix solutions is also NP-hard (see Section 5). As a result, if a given set of r-squares are likely to correspond to multiple haplotype matrices, it is difficult for an attacker to recover a haplotype matrix from a given set of r-squares; and even if she manages to do so, she cannot know if there exists another matrix and how many other matrices may also result in the same r-squares. Since an attacker without exponential sampling power cannot get the distribution of matrices over r-squares, a sufficient condition to safely release a set of r-squares is $\|S\| > \|R^2\|$. By combining Equation 4 and Equation 5, we get the following condition:

$$\frac{2(N-1)}{\log(N+1)-1} > L \tag{10}$$

This condition gives an estimate of the number of individuals needed to ensure that multiple solutions that does not share a sequence in common. Again, we need to consider a human genetic model while analyzing the solution space as we did in section 4, which can lead to a more conservative condition.

F Analysis algorithm

We devised an *iterative algorithm* to approximate the probability distribution over all 2^L sequences. Given an MC model and a threshold θ, in each iteration step (from $i = 1$ to L), the probability of a subsequence $P(X_1X_2...X_i)$ can be computed from the probability of its $i-1$ prefix $P(X_1X_2...X_{i-1})$, i.e., $P(X_1X_2...X_i) = P(X_1X_2...X_{i-1}) \cdot P^{i-1}(X_i|X_{i-1})$. To accommodate our computation into a reasonable memory usage, in each step i, we group all 2^i sequences into $2T$ (by default, we set $T = 2^{12} = 4096$) bins based on their probabilities, where the bins $B_1^0 \cdots B_T^0$ are used for the sequences with a major allele at i ($X_i = 0$), and the bins $B_1^1 \cdots B_T^1$ are used for the sequences with a minor allele at i ($X_i = 1$). Afterwards, we only record the number of sequences and the *average* probability of these sequences in each bin. For instance, the bin B_k^0 ($k = 1, 2, ..., T$) contains all sequences ending with 0 that have a probability within the range of $[p_{min} + \delta(k-1)/T, p_{min} + \delta k/T]$, where $\delta = p_{max} - p_{min}$, and p_{min} and

p_{max} represent the minimum and maximum probability among all sequences of length i, respectively. The number and the average probability of these sequences are recorded. In each step during iteration, we limit the required memory usage to $O(T)$.

The $2T$ bins of initial states can be easily computed because only the probabilities of two sequences (1 and 0) need to be computed, and stored into appropriate bins. Now we consider how to update the $2T$ bins at state $i+1$ from the $2T$ bins at state i. For any sequence $X_1X_2...X_i$ in a bin of B_k^0 (i.e. $X_i = 0$), the probability of the sequence $X_1X_2...X_iX_{i+1}$ can be approximated by $P(X_1X_2...X_iX_{i+1}) = P(B_k^0) \cdot P^i(X_{i+1}|X_i)$, where $P(B_k^0)$ denotes the average probability of sequences in the bin B_k^0 that is computed and stored in the previous step. Accordingly, all sequences in the same bin at state i will receive the same probability as long as they are extended to the same allele type X_{i+1}, and thus one bin at state i should be partitioned into two groups, one for $X_{i+1} = 0$ and one for $X_{i+1} = 1$, each with a different probability at state $i+1$ (thus are possibly assigned to different bins at state $i+1$). To utilize this property, we first compute the probability for each of the in total $4T$ ($= 2T \times 2$) groups of sequences at state $i+1$, then take the minimum and maximum probabilities among them to compute the range of probability for the new bins, and finally assign the $4T$ groups into $2T$ bins based on their probability. Note that the number of sequences in each new bin is simply the total number of sequences in the groups that are assigned to this bin, and average probability of the sequences can be computed by a weighted average of the probabilities of these groups, weighted by the number of sequences in each group. By using the iterative updating algorithm, we can finally obtain the $2T$ bins at the state L, which can be used to estimate the solution space. We first sort the bins in the decreasing order of their average probabilities. Then we count the total number of sequences in the bins and calculate the cumulative probability of the counted sequences by assuming the sequences in the same bin have the same probability as their average The number of sequences when their cumulative probability reaches the threshold θ give the approximate effective solution space. The computational complexity of each iteration step is $O(T)$ and that of the entire algorithm is $O(TL)$.