Online Materials for "Providing Privacy Promises for Aggregate Genomic Data"

Sean Simmons and Bonnie Berger

Contents

1 Online Methods

Overview of Methods

We provide a brief overview of the method here. Those interested in more details should consult later sections.

As in the main paper, assume D is the set of participants in our study. Let d be a given individuals genotype, $x = (x_1, \ldots, x_m)$ the number of times each minor allele occurs in our study population. Note that x is equal to $2n$ times the minor allele frequency in our study. We know that D is drawn from some background population, let us call it B, where B consists of N people, and D consists of n people. For a given individual our privacy measure, which we denote by $\text{PrivMAF}(d, \frac{x}{2n})$, is an upper bound on

$$P\left(d \in \tilde{D} | d \in \tilde{B}, \text{MAF}(\tilde{D}) = \frac{x}{2n}\right)$$

where $\tilde{D}$ and $\tilde{B}$ are chosen from the same distribution as D and B. The PrivMAF score for our study is equal to

$$\text{PrivMAF}(D) = \max_{d \in D} \text{PrivMAF}(d, \text{MAF}(D))$$

In order to calculate PrivMAF we use the equation

$$\text{PrivMAF}(d, \frac{x}{2n}) \approx \frac{1}{1 + \frac{(N-n)P_n(x)}{nP_{n-1}(x-d)}}$$

where

$$P_n(x) = \prod_{i=1}^{m} \binom{2n}{x_i} p_i^{x_i} (1-p_i)^{2n-x_i}$$

This measure can be extended to measure the amount of privacy given by perturbing the MAF (see below).

Basic Model

Before describing the model underlying our results we want to motivate it. Often the size of an underlying population is so large compared to that of the study population that we might as well consider the underlying population to be infinite (consider for example the population of all people of English

ancestry versus the participants in the British Birth Cohort). In practice, however, it might be that we know the study participants are drawn from some smaller subpopulation (for example the British Birth Cohort is drawn from the population of all children born in Britain during a certain week in 1958). This subpopulation is small enough that we can not consider it infinite. Therefore we can think of our study population as being generated by first generating this smaller subpopulation out of an infinite background population, then choosing the study participants out of this smaller population. This is the point of view our model takes, and it is formally described below.

It is worth noting that, breaking with standard notation, we assume all sets are ordered and can have repetitions.

Assume that the genotypes of study participants are drawn from some theoretical infinite population. We have m SNPs, which we label with $1, \ldots, m$, each of which is independent of the others. Let p_i be the minor allele frequency of the ith SNP in our infinite population, and assume our population is in Hardy-Weinberg (H-W) Equilibrium. We first produce a small background population $B = \{b_1, \cdots, b_N\}$ where each $b_j \in \{0, 1, 2\}^m$ (B is the finite set of people who in reality might have participated in the study), and where each member of the population is generated independently of the others. Our study population, denoted $D = \{z_1, \cdots, z_n\}$, is a population of size n produced from the background population by choosing n members of B uniformly at random with no repetitions (note, since B can have repetitions in it, it is possible to have $z_i = z_j$ even if $i \neq j$. This is because it is possible to have $k \neq l$ so that $b_k = b_l$.). It is worth noting that the marginal probability distribution on D is exactly the same as the probability distribution we would get by generating D directly from the infinite population.

PrivMAF

Let $\text{MAF}_i(D)$ be the minor allele frequency of the ith SNP in our population D. We want to release $\frac{x}{2n} = \text{MAF}(D) = (\text{MAF}_1(D), \ldots, \text{MAF}_m(D))$ (where $x_i = \sum_{d \in D} d_i$ is the number of times the minor allele occurs at SNP i in our study population). To simplify notation let $x(D) = 2n\text{MAF}(D)$. We want some kind of measure of how much privacy is lost by each study participant after releasing $\text{MAF}(D)$. We achieve this goal by measuring the probability

that an individual participated in the study given the data released. For a given individual d we want to calculate how likely it is under our model that d is in D given $x(D)$. Note that (in practice) we know that $d \in B$ (that is to say if an adversary is trying to figure out if $d \in D$ they already know $d \in B$), so what we want to calculate is the probability that d is in D conditional on d being in B and on x equaling $x(D)$. More formally, we want to consider:

$$P(d \in \tilde{D} | d \in \tilde{B}, x(\tilde{D}) = x)$$

where $\tilde{D}$ and $\tilde{B}$ have the same distribution as D and B. We would like to devise a formula to calculate an upper bound on this probability. First we need to build a few tools.

We can write $D = \{z_1, \ldots, z_n\}$ and $D = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$.

Let $\tilde{B} - \tilde{D}$ be the set of all people in $\tilde{B}$ who are not in $\tilde{D}$. Note $\tilde{B} - \tilde{D}$ and $\tilde{D}$ are independent random variables, so

$$P(d \in \tilde{B}, d \notin \tilde{D} | x(\tilde{D}) = x) = P(d \in \tilde{B} - \tilde{D} | x(\tilde{D}) = x) P(d \notin \tilde{D} | x(\tilde{D}) = x)$$

$$= P(d \in \tilde{B} - \tilde{D}) P(d \notin \tilde{D} | x(\tilde{D}) = x) = P(d \in \tilde{B} - \tilde{D})(1 - P(d \in \tilde{D} | x(\tilde{D}) = x))$$

We also see, since $d \in \tilde{D}$ implies $d \in \tilde{B}$, that

$$P(d \in \tilde{D}, d \in \tilde{B} | x(\tilde{D}) = x) = P(d \in \tilde{D} | x(\tilde{D}) = x)$$

Using Bayes' rule and some algebra we see that

$$P(d \in \tilde{D} | d \in \tilde{B}, x(\tilde{D}) = x) = \frac{P(d \in \tilde{D}, d \in \tilde{B} | x(\tilde{D}) = x)}{P(d \in \tilde{D} | x(\tilde{D}) = x) + P(d \in \tilde{B}, d \notin \tilde{D} | x(\tilde{D}) = x)}$$

$$= \frac{P(d \in \tilde{D} | x(\tilde{D}) = x)}{P(d \in \tilde{D} | x(\tilde{D}) = x) + P(d \in \tilde{B} - \tilde{D})(1 - P(d \in \tilde{D} | x(\tilde{D}) = x))}$$

$$= \frac{1}{1 + P(d \in \tilde{B} - \tilde{D})(\frac{1}{P(d \in \tilde{D} | x(\tilde{D}) = x)} - 1)} \tag{1}$$

The next step is to consider $P(d \in \tilde{D} | x(\tilde{D}) = x)$. This equals

$$1 - P(d \notin \tilde{D} | x(\tilde{D}) = x) = 1 - \prod_{i=1}^{n} P(d \neq \tilde{z}_i | x(\tilde{D}) = x) = 1 - (1 - P(d = \tilde{z}_1 | x(\tilde{D}) = x))^n$$

Then note that

$$P(d = \tilde{z}_1 | x(D) = x) = \frac{P(d = \tilde{z}_1, x(\tilde{D}) = x)}{P(x(\tilde{D}) = x)}$$

$$= \frac{P(d = \tilde{z}_1, x(\tilde{z}_2, \cdots, \tilde{z}_n) = x - d)}{P(x(\tilde{D}) = x)}$$

$$= \frac{P(d = \tilde{z}_1) P(x(\tilde{z}_2, \cdots, \tilde{z}_n) = x - d)}{P(x(\tilde{D}) = x)} \tag{2}$$

Let $P_n(x) = P(x(\tilde{D}) = x)$; then equation 2 equals

$$= P(d = \tilde{z}_1) \frac{P_{n-1}(x - d)}{P_n(x)}$$

Substituting this in to equation 1 we get that

$$P(d \in \tilde{D} | d \in \tilde{B}, x(\tilde{D}) = x) = \frac{1}{1 + \left(\frac{P(d \in \tilde{B} - \tilde{D})}{1 - \left(1 - P(d = \tilde{z}_1) \frac{P_{n-1}(x-d)}{P_n(\tilde{D})}\right)^n} \right) - P(\tilde{D} \in \tilde{B} - \tilde{D})}$$

Using the fact that $(1 - z)^n \geq 1 - nz$ when $0 \leq z \leq 1$ (this follows from the inclusion exclusion principle) we get that

$$\leq \frac{1}{1 - P(d \in \tilde{B} - \tilde{D}) + \frac{P(d \in \tilde{B} - \tilde{D})(P_n(x))}{nP(d = \tilde{z}_1) P_{n-1}(x-d)}} = \text{PrivMAF}(d, \text{MAF}(D))$$

It is worth mentioning that the above upper bound is likely to be fairly tight, since $z = P(d = \tilde{z}_1 | x(\tilde{D}) = x)$, which in practice is likely to be very small (especially when the data set is anywhere near being safe to release, since $P(d = \tilde{z}_1 | x(\tilde{D}) = x) \leq P(d = \tilde{z}_1 | d \in \tilde{B}, x(\tilde{D}) = x)$).

This quantity, $\text{PrivMAF}(d, \text{MAF}(D))$, is our measure of privacy.

Note that for realistic choices of n, N, p and m we get that $P(d \in \tilde{B} - \tilde{D})$ is approximately equal to $(N - n)P(d = \tilde{z}_1)$ and that $P(d \in \tilde{B} - \tilde{D}) << 1$, so $1 - P(d \in \tilde{B} - \tilde{D}) \approx 1$. Plugging this in we get the measure

$$\text{PrivMAF}(d, \text{MAF}(D)) \approx \frac{1}{1 + \frac{(N-n)P_n(x)}{nP_{n-1}(x-d)}}$$

which is what we use in practice. Moreover, we see that

$$P_n(x) = \prod_{i=1}^{m} \binom{2n}{x_i} p_i^{x_i} (1-p_i)^{2n-x_i}$$

This allows us to calculate $\text{PrivMAF}(d, \text{MAF}(D))$ easily.

PrivMAF allows us to determine how much privacy is lost by a particular individual. What we want is the total privacy loss by releasing a study. It makes sense to look at maximum loss to any individual in our study, which is to say $\max_{d \in D} \text{PrivMAF}(d, \text{MAF}(D))$. We call this quantity PrivMAF. If PrivMAF is bounded above by α then, for any participant $d \in D$, an adversary can be at most α percent confident that d actually participated in the study, which is the privacy guarantee we want.

Naively it seems like calculating PrivMAF for m SNPs and n individuals has complexity $O(mn^2)$. By using the cancellation in $\frac{P_n(x)}{P_{n-1}(x-s)}$ it only ends up taking $O(mn)$ time, which is asymptotically optimal.

PrivMAF for Data with Noise Added

Note that the above framework can be generalized to measure the privacy loss present in releasing noisy version of $\text{MAF}(D)$. In particular, let η be some random variable. Then we can let $MAF_j^{\eta}(D) = MAF_j(D) + \frac{\eta_j}{2n}$, where $\eta_1, \ldots, \eta_m$ are iid random variables distributed as η. Then we want to measure how well $\text{MAF}^{\eta}(D)$ preserves privacy. As above, we are interested in $P(d \in \tilde{D} | \text{MAF}^{\eta}(\tilde{D}) = \text{MAF}^{\eta}(D), d \in \tilde{B})$. The same derivation used in the previous section implies that this probability is upper bounded by:

$$\text{PrivMAF}^{\eta}(d, \text{MAF}^{\eta}(D)) = \frac{1}{1 - P(d \in \tilde{B} - \tilde{D}) + \frac{P(d \in \tilde{B} - \tilde{D})(P_n^{\eta}(\text{MAF}^{\eta}(D)))}{nP(d = \tilde{z}_1) P_{n-1}^{\eta}(\text{MAF}^{\eta}(D) - d)}}$$

where

$$P_n^{\eta}(v) = \prod_{j=1}^{m} \sum_{i=0}^{2n} \binom{2n}{i} p_j^i (1-p_j)^{2n-i} P(\eta = 2nv_j - i)$$

Note that the same approximations used in the previous section apply here. In this paper we will let $P(\eta = i)$ be chosen proportional to $e^{-\epsilon|i|}$,

where i is an integer and ϵ is a user chosen privacy parameter (relating to ϵ-differential privacy guarantees). We can then let

$$\text{MAF}^{\epsilon} = \text{MAF}^{\eta}$$

and

$$\text{PrivMAF}^{\epsilon} = \text{PrivMAF}^{\eta}$$

PrivMAF for Data with Truncation

Similarly we can consider the gain in privacy we get by rounding our MAF. More specifically, consider $k > 1$, then if $MAF_j(D) = \frac{x_j}{2n}$, we let $MAF_j^{\text{trunc}(k)}(D)$ be the result of truncating each entry in $\text{MAF}(D)$ after k decimal digits. More formally

$$MAF_j^{\text{trunc}(k)}(D) = \frac{\lfloor MAF_j(D) * 10^k \rfloor}{10^k}$$

In the below we let $v = \text{MAF}^{\text{trunc}(k)}(D)$ to make the equations more readable. In order to measure privacy we want to calculate $P(d \in \tilde{D} | \text{MAF}^{\text{trunc}(k)}(\tilde{D}) = v, d \in \tilde{B})$. As above, we can upper bound this by

$$\frac{1}{1 - P(d \in \tilde{B} - \tilde{D}) + \frac{P(d \in \tilde{B} - \tilde{D})}{nP(d = \tilde{z}_1)} \frac{P(\text{MAF}^{\text{trunc}(k)}(\tilde{D}) = v)}{P(\text{MAF}^{\text{trunc}(k)}(\tilde{D}) = v | d = \tilde{z}_1)}} = \text{PrivMAF}^{\text{trunc}(k)}(d, \text{MAF}^{\text{trunc}(k)}(D))$$

Note

$$P(\text{MAF}^{\text{trunc}(k)}(\tilde{D}) = v) = \prod_{j=1}^{m} P(MAF_j^{\text{trunc}(k)}(\tilde{D}) = v_j)$$

and

$$P(\text{MAF}^{\text{trunc}(k)}(\tilde{D}) = v | d = \tilde{z}_1) = \prod_{j=1}^{m} P(MAF_j^{\text{trunc}(k)}(\tilde{D}) = v_j | d = \tilde{z}_1)$$

If $S_k(v_j) = \{x | \frac{x}{2n} \text{ truncates to } v_j\}$, then

$$P(MAF_j^{\text{trunc}(k)}(\tilde{D}) = v_j) = \sum_{i \in S_k(v_j)} \binom{2n}{i} p_j^i (1 - p_j)^{2n-i}$$

and

$$P(MAF_j^{\text{trunc}(k)}(\tilde{D}) = v_j | d = \tilde{z}_1) = \sum_{i \in S_k(v_j)} \binom{2n-2}{i-d_j} p_j^i (1-p_j)^{2n-i+d_j-2}$$

This allows us to calculate $\text{PrivMAF}^{\text{trunc}(k)}(d, MAF_j^{\text{trunc}(k)}(D))$, just as we wanted.

Comparison to previous approaches

Our approach is the first to give privacy guarantees for all individuals in a study. The method endorsed by Sankararaman et al. [3] provides guarantees of a sort– since the log likelihood test gives the best power for a given false positive ratio, it ensures that the power of any test can not be too big. The problem with their guarantee is that it is an aggregate guarantee, and does not ensure the safety of all participants. Our approach, on the other hand, does ensure privacy for all involved. Our method also takes into account the size of the pool from which our study is drawn, something the likelihood approach does not take into account but which is important in measuring privacy. The PPV approach suggested by Craig et al. [4] does take the background population size into account, but again does not come with any privacy guarantees that hold for all participants. We also believe that our method gives a more intuitive measure of privacy than previous ones (though of course this is subjective). One might argue that the difference between the worse case and average case privacy loss are not that different, but our experiments do not seem to support this claim (see, for example, **Fig S1** that compares the value of $\max_{d \in D} \text{PrivMAF}(d, \text{MAF}(D))$ for a random choice of D with the worse case value.)

Zhou, et al. [5] have also presented work with strong privacy guarantees; however, they examined the frequency and likelihood of pairs of alleles rather than MAF. Moreover, they give guarantees of a combinatorial nature (using k-anonymity), where as ours are probabilistic in nature.

A Release Mechanism: Allele Leakage Guarantee Test

Motivation

As mentioned in the paper, we want to use the above measure to decide if it is safe to release MAF(D) (Note that we could do something similar for perturbed MAF, but do not do so here). Assume we want to bound the probability of an adversary figuring out if someone took part in the study to be at most α, where $0 < \alpha < 1$. A first guess at how to do this might be to look at PrivMAF, and release if and only if it is at most α.

Though in practice this approach seems to work well, in theory we can have trouble. The problem with this approach is that the decision of whether or not to release gives away a little information about D and thus destroys our probability guarantees (to understand why this is note that deciding to release if and only if PrivMAF is less than α means that any data being released gives away two pieces of information, namely the value of MAF and the fact that PrivMAF is less than α. If PrivMAF is greater than α with non-negligible probability (say 50 percent probability or so) this extra bit of information can actually be very informative).

An obvious fix is to release if and only if $\max_{d \in \{0,1,2\}^m} \text{PrivMAF}(d, \text{MAF}(D)) \leq \alpha$. In this case the decision of whether or not to release gives no more information than outputting MAF(D) by itself. It turns out that this quantity is easy to calculate, and gives us the security guarantee we want. Unfortunately it is also overkill: the worst-case behavior is often much worse than the average case, so this policy is likely to tell us a data set is not safe to release even when it is (see **Fig S1**).

This leads us to propose another solution. For any choice of β we can define

$$P_\beta = P(\max_{d \in \tilde{D}} \text{PrivMAF}(d, \text{MAF}(\tilde{D})) \leq \beta | x(\tilde{D}) = x)$$

where the probability is taken over the choice of $\tilde{D}$. Choose β so that

$$\alpha \geq \frac{1}{1 + \frac{P_\beta}{\beta} - P_\beta - \max_{d \in \{0,1,2\}^m} P(d \in \tilde{B} - \tilde{D})}$$

and release the data if and only if PrivMAF(D) is less than or equal to β. This release test is what is referred to as the Allele Leakage Guarantee Test (ALGT) in the paper. We can show that ALGT gives us the privacy we

require without too much overkill. On the other hand it is much slower than the above methods, since calculating P_β is slow (described below).

Derivation

ALGT tells us that, given both MAF(D) and the knowledge leaked by the decision to release, then from the adversaries view the probability that $d \in D$ is at most α for any choice of $d \in D$. More formally:

Theorem 1. *Choose β as above. Then, if PrivMAF(D) $\leq \beta$, for any choice of $d \in D$ we get that*

$$P\left(d \in \tilde{D} | d \in \tilde{B}, x(\tilde{D}) = x, \max_{\tilde{d} \in \tilde{D}} PrivMAF(\tilde{d}, MAF(\tilde{D})) \leq \beta\right) \leq \alpha$$

Proof. The proof basically comes down to repeated applications of the definition of conditional probability, independence, and Bayes rule. Let R be the event that $\max_{i=1,\ldots,n} \text{PrivMAF}(\tilde{z}_i, \text{MAF}(\tilde{D})) \leq \beta$. Then

$$P(d \in \tilde{D} | d \in \tilde{B}, x(\tilde{D}) = x, \max_{\tilde{d} \in D} \text{PrivMAF}(\tilde{d}, \text{MAF}(\tilde{D})) \leq \beta) = \frac{P(d \in \tilde{D} | x(\tilde{D}) = x, R)}{P(d \in \tilde{B} | x(\tilde{D}) = x, R)}$$

$$= \frac{P(d \in \tilde{D} | x(\tilde{D}) = x, R)}{P(d \in \tilde{D} | x(\tilde{D}) = x, R) + P(d \in \tilde{B} - \tilde{D})(1 - P(d \in \tilde{D} | x(\tilde{D}) = x, R))}$$

$$= \frac{1}{1 + \frac{P(d \in \tilde{B} - \tilde{D})}{P(d \in \tilde{D} | x(\tilde{D}) = x, R)} - P(d \in \tilde{B} - \tilde{D})} \tag{3}$$

To simplify this note that

$$\frac{P(d \in \tilde{D} | x(\tilde{D}) = x, R)}{P(d \in \tilde{B} - \tilde{D})} = \frac{P(R, d \in \tilde{D} | x(\tilde{D}) = x)}{P(d \in \tilde{B} - \tilde{D}) P_\beta}$$

$$\leq \frac{P(d \in \tilde{D} | x(\tilde{D}) = x)}{P(d \in \tilde{B} - \tilde{D}) P_\beta} \leq \frac{\text{PrivMAF}(d, \text{MAF}(D)) P(d \in \tilde{B} | x(\tilde{D}) = x)}{P(d \in \tilde{B} - \tilde{D}) P_\beta}$$

To simplifying this we look at $\frac{P(d \in \tilde{B} | x(\tilde{D}) = x)}{P(d \in \tilde{B} - \tilde{D})}$, which we see equals

$$= \frac{P(d \in \tilde{D}|x(\tilde{D}) = x) + P(d \in \tilde{B} - \tilde{D})(1 - P(d \in \tilde{D}|x(\tilde{D}) = x))}{P(d \in \tilde{B} - \tilde{D})}$$

$$= 1 + \frac{P(d \in \tilde{D}|x(\tilde{D}) = x)(1 - P(d \in \tilde{B} - \tilde{D}))}{P(d \in \tilde{B} - \tilde{D})}$$

Using the fact that $P(d \in \tilde{B} - \tilde{D}) = P(d \in \tilde{B} - \tilde{D}|x(\tilde{D}) = x)$ this becomes

$$= 1 + \frac{P(d \in \tilde{D}|x(\tilde{D}) = x)(1 - P(d \in \tilde{B} - \tilde{D}))}{P(d \in \tilde{B}|x(\tilde{D}) = x) - P(d \in \tilde{D}|x(\tilde{D}) = x) + P(d \in \tilde{D}|x(\tilde{D}) = x)P(d \in \tilde{B} - \tilde{D})}$$

$$= 1 + \frac{P(d \in \tilde{D}|d \in \tilde{B}, x(\tilde{D}) = x)(1 - P(d \in \tilde{B} - \tilde{D}))}{1 - P(d \in \tilde{D}|d \in \tilde{B}, x(\tilde{D}) = x) + P(d \in \tilde{D}|d \in \tilde{B}, x(\tilde{D}) = x)P(d \in \tilde{B} - \tilde{D})}$$

$$\leq 1 + \frac{\text{PrivMAF}(d, \text{MAF}(D))(1 - P(d \in \tilde{B} - \tilde{D}))}{1 - \text{PrivMAF}(d, \text{MAF}(D)) + \text{PrivMAF}(d, \text{MAF}(D))P(d \in \tilde{B} - \tilde{D})}$$

$$= 1 + \frac{1}{\frac{1}{(1 - P(d \in \tilde{B} - \tilde{D}))\text{PrivMAF}(d, \text{MAF}(D))} - 1} \leq 1 + \frac{1}{\frac{1}{\text{PrivMAF}(d, \text{MAF}(D))} - 1} \leq 1 + \frac{1}{\frac{1}{\beta} - 1}$$

Substituting this in to equation 3 results we get

$$\frac{P(d \in \tilde{D}|x(\tilde{D}) = x, R)}{P(d \in \tilde{B} - \tilde{D})} = \frac{\text{PrivMAF}(d, \text{MAF}(D))}{P_\beta}\left(1 + \frac{1}{\frac{1}{(1 - P(d \in \tilde{B} - \tilde{D}))\text{PrivMAF}(d, \text{MAF}(D))} - 1}\right)$$

$$\leq \frac{\text{PrivMAF}(d, \text{MAF}(D))}{P_\beta}\left(1 + \frac{1}{\frac{1}{\text{PrivMAF}(d, \text{MAF}(D))} - 1}\right) \leq \frac{\text{PrivMAF}(d, \text{MAF}(D))}{P_\beta}\left(1 + \frac{1}{\frac{1}{\beta} - 1}\right)$$

Putting it all together we see that

$$P(d \in \tilde{D}|d \in \tilde{B}, x(\tilde{D}) = x, R) \leq \frac{1}{1 + \frac{1}{\frac{\text{PrivMAF}(d, \text{MAF}(D))}{P_R}\left(1 + \frac{1}{\frac{1}{\beta} - 1}\right)} - P(d \in \tilde{B} - \tilde{D})}$$

$$= \frac{1}{1 - P(d \in \tilde{B} - \tilde{D}) + \frac{P_\beta}{\beta} - P_\beta}$$

$$\leq \frac{1}{1 + \frac{P_\beta}{\beta} - P_\beta - \max_{d \in \{0,1,2\}^m} P(d \in \tilde{B} - \tilde{D})} \leq \alpha$$

which is what we wanted.

□

Note that, in practice, since $\max_{d \in \{0,1,2\}^m} P(d \in \tilde{B} - \tilde{D}) << 1$, we choose an approximate β such that

$$\alpha \geq \frac{1}{1 + \frac{P_\beta}{\beta} - P_\beta}$$

Comparing β to α

To justify our release test ALGT, we compared the naive threshold, α, to the corrected threshold, β (**Fig. S6**). For larger values of α we see that the two thresholds are fairly close. As α decreases, however, the two quantities start to diverge, with the corrected threshold decreasing much faster than the naive one. Moreover, we see that when α is roughly 0.04, β suddenly drops to around 0 and remains at that level for all smaller α– this behavior is due to the negligible probability that a study population would have an PrivMAF less than .04 given this choice of parameters. This suggests that, in most cases, using α instead of β will not reduce privacy by too much.

Example Application of ALGT

Suppose that the Wellcome Trust wants to publicly release aggregate MAF statistics to facilitate researchers timely data access. Our test can be used to determine whether or not it is safe to do so. Below, we work through an example to better illustrate the details of our method and how it might be used in practice.

As above we start by choosing a set of 1000 participants in the British Birth Cohort–this is our study population. In order to calculate our population parameters we use the remaining 500 individual‘s to estimate the background populations minor allele frequency (in practice these individual‘s can be taken from HapMap or some similar data source). The first step is to estimate N, the size of the population from which our study is drawn. Assume our study is known to be drawn from the population of a city consisting of 100000 people; then we can choose $N = 100000$. Note this might be an overly optimistic choice of N– other information might be known about the study participants that makes N smaller–, but for simplicity we decide to use this value.

We would like to release the MAFs for 200 of the SNPs, but we only want to do so if the probability that $d \in D$ based on the publicly available data is at most 20% (note that 20% is a reasonable choice, as practitioners of

k-anonymity are often advised to use $k = 5$ [1], which can be seen as corresponding to 20% in this context). This selection corresponds to us choosing $\alpha = .2$. For the practitioner using our tool, the choice of the security parameter α, which quantifies the risk of re-identification, is dependent on the context– α should be set based on the level of harm that can result from determining if a given individual is in a study (perhaps using a framework similar to the one proposed by [2] for setting parameters in differential privacy.) Applying our method, we determine a stricter threshold of $\beta = .196$. Note that in this case the adjustment to our threshold α is very small. In general, however, it can be much larger. Applying PrivMAF to our study cohort gives us a score of .177. Since this is less than β we can release the study participants MAF while still preserving privacy at the level required.

Scalability of ALGT

We have presented results on moderate-sized datasets. We have also run our algorithm on larger artificial datasets (with 10,000 individual's and 1000 SNPs) and have found our ALGT implementation still runs in a reasonable amount of time, completing in just over 8 hours on a single core (Methods). Although our current implementation runs on a single core, the PrivMAF framework permits parallelization of Monte Carlo sampling, the major computational bottleneck in our pipeline, i.e. computing β, and thus is able to benefit from any parallel or distributed computing system. As dataset sizes grow, we expect to be able to keep pace by computing the PrivMAF statistic more efficiently.

Choosing which SNPs to release

Often one would like to release the maximum number of minor allele frequencies that still gives us our privacy guarantee. Unfortunately, doing so can give away a lot of information about our participants– we need to know the SNPs we want to release ahead of time. Therefore we suggest choosing the m SNPs we want to release by choosing a set of SNP so that P_β is large (almost 1). This approach preserves the privacy guarantee and makes it very likely that the user will get to release their data set to the public (the probability of not being able to release it is $1 - P_\beta$). Alternatively we can use the measure $\max_{d \in \{0,1,2\}^m} \text{PrivMAF}(d, \text{MAF}(D))$ to decide if we want to release our SNPs or not, in which case we can choose the largest set of SNPs possible without

giving up our privacy guarantee. It is plausible that there are other ways of picking m as well.

Estimating P_β

Unfortunately, $P_\beta = P(\text{PrivMAF}(D) \leq \beta | x(D) = x)$ is not so easy to calculate. We use a Monte Carlo type approach to calculate it. More precisely we sample D conditional on $x(D) = x$, then estimate P_β as being the percentage of the D we generated for which $\max_{d \in D} \text{PrivMAF}(d, \text{MAF}(D)) \leq \beta$.

This approach requires us to be able to sample D such that $x(D) = x$. In order to do this consider $t_i = \#\{j | z_{j,i} = 2\}$, where $z_{j,i}$ is the genotype of z_j at SNP i. Then the probability that $t_i = t$ is proportional to

$$\binom{n}{t}\binom{n-t}{n+t-x_i} p_i^{2t} (2p_i(1-p_i))^{x_i - 2t} (1-p_i)^{2(n+t-x_i)}$$

where we hold to the convention that $\binom{n}{m} = 0$ if $n \leq 0$, $m < 0$ or $n < m$. This allows us to sample from t_i. Knowing t_i we can then calculate the number of j so that $z_{j,i} = 1$ (namely $x_i - 2t_i$) and the number that equal 0 (namely $n + t_i - x_i$). We can then randomly choose t_i individuals to have $z_{j,i} = 2$, and similarly for $z_{j,i} = 1$ and $z_{j,i} = 0$. Repeating this process for all of the SNPs gives us a random sample of D conditional on $x(D) = x$.

Of course Monte Carlo estimation is often very slow. What alternatives do we have? One is to note that, if $M = \log\left(\frac{n}{N-n}\left(\frac{1}{\text{PrivMAF}(z_1, \frac{x}{2n})} - 1\right)\right)$ then (conditional on $x(D) = x$) as m goes to infinity we get that (under reasonable assumptions, such as a MAF $\geq .05$, n fixed)

$$\frac{M - EM}{\sqrt{var(M)}} \Rightarrow \chi$$

where EM is the expected value of M, $var(M)$ is its variance (both of which we can calculate), and χ is a unit normal centered at 0. This result follows from considering the distribution, Q, on the pairs x_i, p_i. Using the Central Limit Theorem, it is straightforward to show the result holds when Q has finite support. One can then use a limiting argument to show it holds for more general Q (we do not include the detail here). This fact gives us a means of estimating $P(\text{PrivMAF}(d_1, \text{MAF}(D)) \leq \beta | x(D) = x)$, which can then be used to estimate P_β

Unfortunately this is only an asymptotic bound, and experiments show that it often gives poor estimates in practice, so we have chosen not to use it in practice. It can be hoped, however, that more robust approximations are possible to speed up this calculation.

Approximating β

Calculating β can be quite time consuming, so one might be tempted to try to avoid calculating β. One way to do this could be to use α instead of β. Experiments on simulated datasets show that there is some β_0 so that if α is above β_0 then β is about equal to α, while below β_0 we see β quickly decays to 0 (this can be seen, for example, in **Fig S6**). This implies that, if PrivMAF is significantly below α (where we do not attempt to define significantly below here), then we should expect β to be close to α, so PrivMAF should be below β as well. This is a heuristic, but this line of reasoning seems like it could lead to something more reliable– more work is needed to know for sure.

Estimating the parameters

The above model require estimates of the p_i. How are they estimated? The straightforward method is to take another collection of individuals (our reference population) drawn from the same background population as our study participants. The minor allele frequencies of this population can then serve as an estimate of the minor allele frequencies for the background population. Alternatively, we can estimate the p_i parameters from the union of this collection of individuals with the study participants, a method advocated by some previous papers.

An alternative approach is to use Bayesian methods to place a prior on p_i, which can then be updated based on the data in the outside population. We can then use this posterior probability on p_i to estimate $P(x_i(D) = x_i)$. In our results we used the naive approach, though arguments can be made for the other two.

The other parameter that one must consider is N, the size of the background population. This depends a lot on the context, and giving a realistic estimate of it is critical. In most applications the background population from which the study is drawn is fairly obvious. That being said, one needs to be careful of any other information released in the paper about participants– just listing a few facts about the participants can greatly reduce N, greatly

reducing the bounds on privacy guarantees (since the probability of a privacy compromise is roughly inversely proportional to $N - n$).

Changing the Assumptions

The above model makes a few assumptions (assumptions that are present in all previous work that we are aware of, with one exception [5]). In particular it assumes that there is no linkage disequilibrium (LD) (which is to say that the SNPs are independently sampled), that the genotypes of individuals are independent of one another (that there are no relatives, population stratification, etc. in the population), and that the background population is in Hardy-Weinberg Equilibrium (H-W Equilibrium). The assumption that genotypes of different individuals are independent from one another is difficult to remove, and we do not consider it here. We can, however, remove either the assumption of H-W Equilibrium or of SNPs being independent.

First consider the case of H-W Equilibrium. Let us consider the ith SNP, and let p_i be the minor allele frequency. We also let $p_{0,i}$, $p_{1,i}$ and $p_{2,i}$ be the probability of us having zero, one, or two copies of the minor allele respectively. Assuming the population is in H-W equilibrium is the same as assuming that $p_{0,i} = (1 - p_i)^2$, $p_{1,i} = 2p_i(1 - p_i)$, and $p_{2,i} = p_i^2$. Dropping this assumption, we see that all of the calculations above still hold, except we get that

$$P(x_i(\tilde{D}) = x_i) = \sum_{c=0}^{\lfloor \frac{x_i}{2} \rfloor} \binom{n}{c} \binom{n-c}{x_i - 2c} p_{0,i}^{n-x_i+c} p_{1,i}^{x_i-2c} p_{2,i}^{c}$$

where we use the convention that $\binom{n}{m} = 0$ when $m < 0$. This allows us to remove the assumption of H-W Equilibrium. Unfortunately there are two problems with this approach. The first is statistical– instead of having to just estimate one parameter per SNP (p_i), we have to estimate two ($p_{0,i}$ and $p_{1,i}$, since $p_{2,i}$ can be calculated from the other two). The other problem is that calculating $P(x_i(D) = x_i)$ suddenly becomes more computationally intensive, so much so that it is prohibitive for large data sets.

In order to allow us to drop the assumption of no LD we can model the genome as a Markov model (you could also use an hidden Markov model instead which allows for more complex relationships, but for simplicity sake we will only talk about Markov models since the generalization to HMM is straightforward). In such a model the state of a given SNP only depends on

the state of the previous SNP. To specify such a model we need to specify the probability distribution of the first SNP, and for each subsequent SNP we need to specify its distribution conditional on the previous SNP. It is then straightforward to modify our framework to deal with this model. As above, however, this requires us to estimate lots of parameters and also is much more time consuming; thus it is not likely to be useful in practice.

Note that the above method assumes our sample is drawn uniformly from the background population, which has various consequences. Strictly speaking the above method should not, for example, be naively applied to case-control studies, only for studies of quantitative traits (this is a weakness shared with previous approaches). This is for two reasons: first, of all our measure does not take into account the fact that most case control studies only release data about SNPs that are highly differentiated between the case and control groups; second, it does not consider the fact that the case population is not drawn from the general background population, but is instead drawn from a population of people with the disease. We can overcome these drawbacks if we have information about the minor allele frequencies of the SNPs we are looking at in the disease population ahead of time. This is likely to be possible in follow up studies, but not in a study that looks at a given disease for the first time.

It should also be mentioned that, in practice, we believe our ALGT bounds will still hold even if we apply them naively– since the disease population is further from average member of our background population than most people we expect our method to overestimate how much information is leaked by releasing the data. This implies that one can use ALGT even in a case-control framework.

Reidentification Using PrivMAF

Thus far we have presented PrivMAF as a means of helping ensure participant privacy. As it turns out, PrivMAF can also be used in exactly the opposite way, as a means of compromising subjects privacy. To do this choose some threshold γ. For a given genotype d we predict $d \in D$ if and only if $\text{PrivMAF}(d, \text{MAF}(D)) > \gamma$. Used in this way, our approach performs comparably to previous approaches; we plot the ROC curve of the likelihood ratio test [3] as well as the ROC curve obtained by using our test statistic (see **Fig S2**). We see that both methods perform similarly. Since it is known that the likelihood ratio test gives the highest power for a given false positive rate of

any test, this curve suggest that our privacy measure is doing a good job in terms of measuring how much privacy is lost in a given dataset by releasing the minor allele frequencies.

Note that we can also use this as a reidentification on perturbed data, using the perturbed PrivMAF. The results of this analysis are shown in **Fig S3** for truncated data and **Fig S4** for data with noise added.

References

[1] G. Loukides, A. Gkoulalas-Divanis, B. Malin, *Anonymization of electronic medical records for validating genome-wide association studies*, PNAS, 107 (2010), pp 78987903.

[2] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. Pierce, A.Roth, *Differential privacy: an economic method for choosing epsilon*, CoRR (2014), abs/1402.3329.

[3] S. Sankararaman, G. Obozinski, M. Jordan, E. Halperin, *Genomic privacy and the limits of individual detection in a pool*, Na.Genet, 41 (2009), pp 965-967.

[4] D. Craig, R. Goor, Z. Wang, J. Paschall, J. Ostell, M. Feolo, S. Sherry, T. Manolio, *Assessing and mitigating risk when sharing aggregate genetic variant data*, Nat Rev Genet, 12 (2011),pp 730-736.

[5] X. Zhou, B. Peng, Y. Li, Y. Chen, H. Tang, X. Wang, *To release or not to release: evaluating information leaks in aggregate human-genome data*, ESORICS 2011, pp 607-627.

2 Online Figures

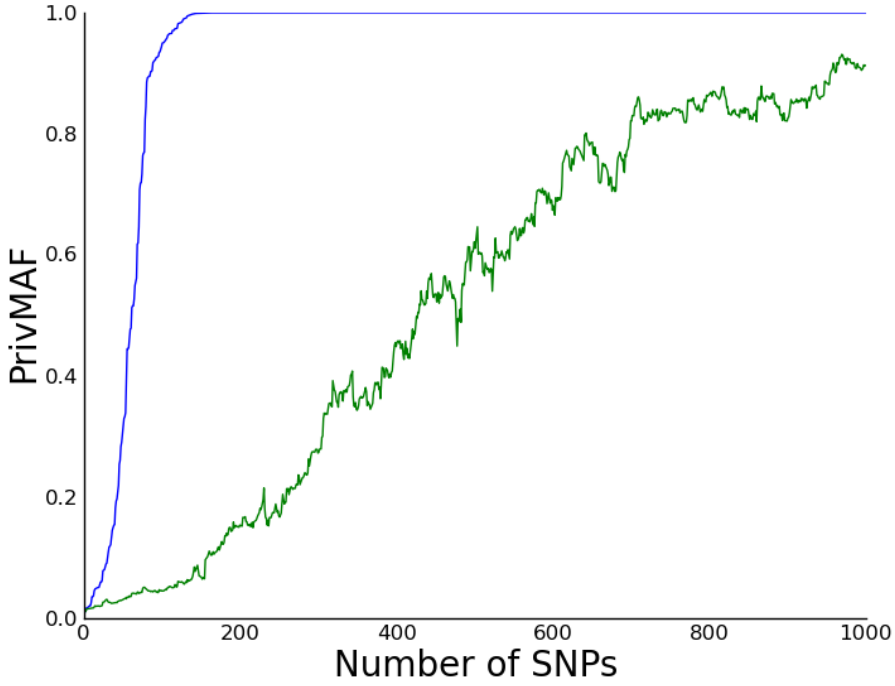

Figure S1: Worst Case Versus Average Case PrivMAF. Graph of the number of SNPs, denoted m, versus PrivMAF. The green curve is the PrivMAF for a set of $n = 1,000$ randomly chosen participants in the British Birth Cohort, while the blue curve is the worst case PrivMAF over all possible choices of our study population conditional on $\text{MAF}(D)$ (in other words $\max_{d \in \{0,1,2\}^m} \text{PrivMAF}(d, \text{MAF}(D))$).

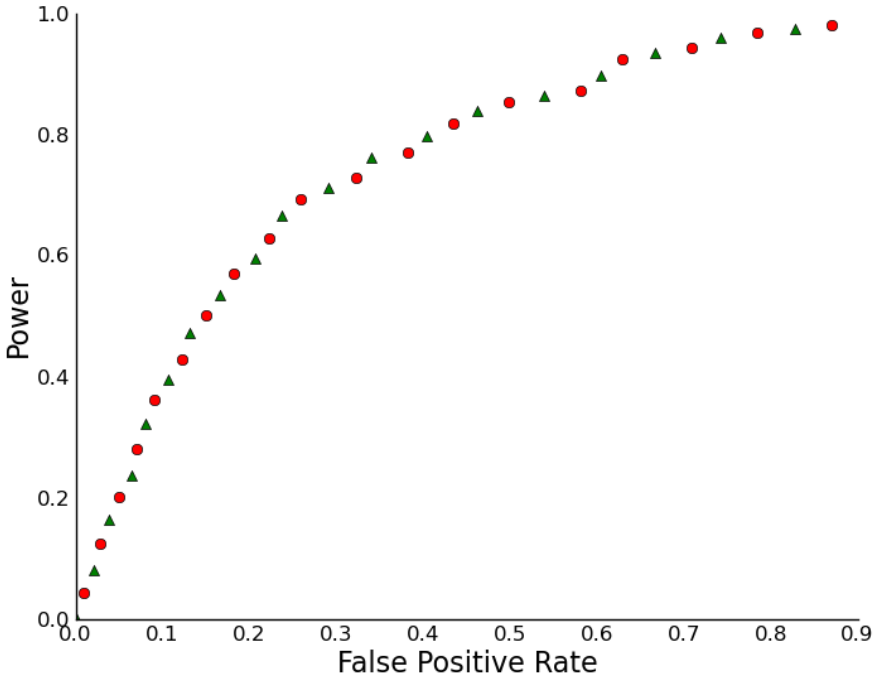

Figure S2: ROC Curves of PrivMAF and Likelihood Ratio. ROC curves obtained using PrivMAF (green triangles) and the likelihood ratio method (red circles) to reidentify individuals in the WTCCC British birth cohort with n=1,000 study participants and 1,000 SNPs.

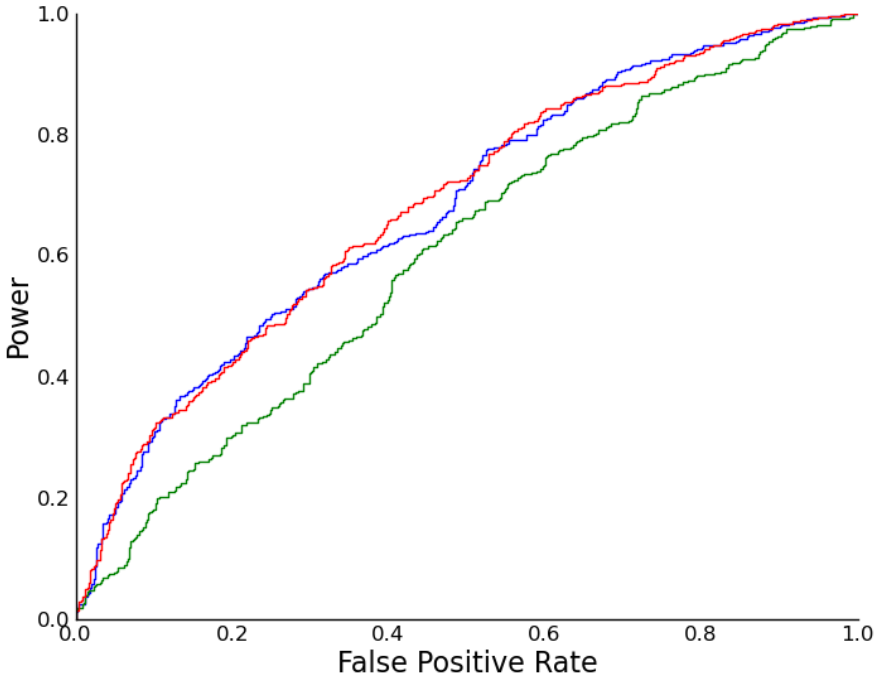

Figure S3: ROC Curves of PrivMAF with Truncation. ROC curves obtained using PrivMAF for reidentification of unperturbed data (in red, AUC=.686), data truncated after two decimal digits (aka $k = 2$, in blue, AUC=.682), and data truncated after one decimal digit (aka $k = 1$, in green, AUC=.605). We see that truncation can greatly decrease the effectiveness of reidentification. Note that the ROC of the unperturbed data here is different from that in the previous figure. This is because we used a different random division of our data in each case.

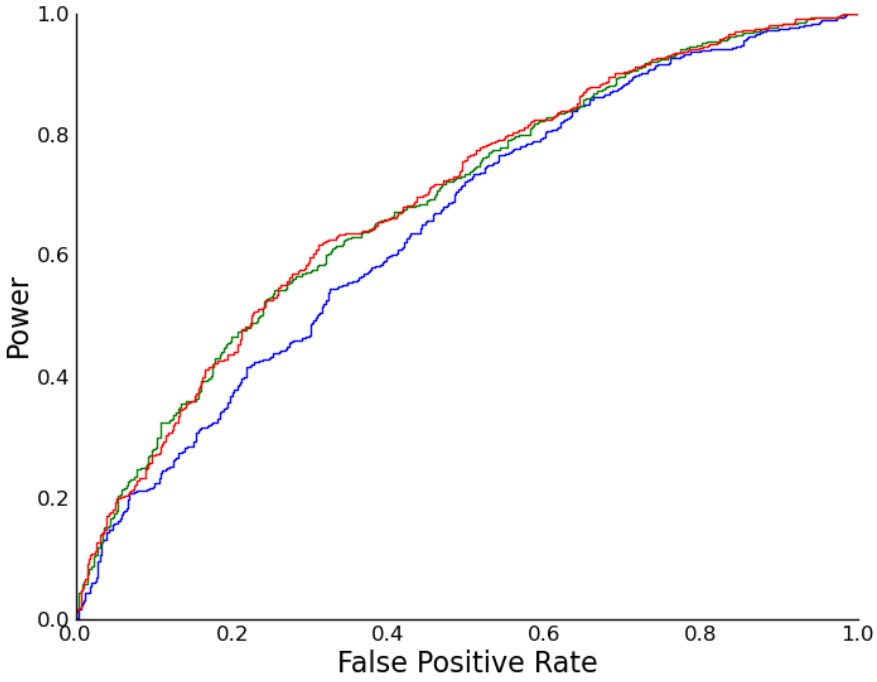

Figure S4: ROC Curves of PrivMAF with noisy data. ROC curves obtained using PrivMAF for reidentification of unperturbed data (in red, AUC=.696), with noise corresponding to $\epsilon = .5$ (in green, AUC=.693), and with $\epsilon = .1$ (in blue, AUC=.656). We see that adding noise can decrease the effectiveness of reidentification. Note that the ROC of the unperturbed data here is different from that in the previous figures. This is because we used a different random division of our data in each case.

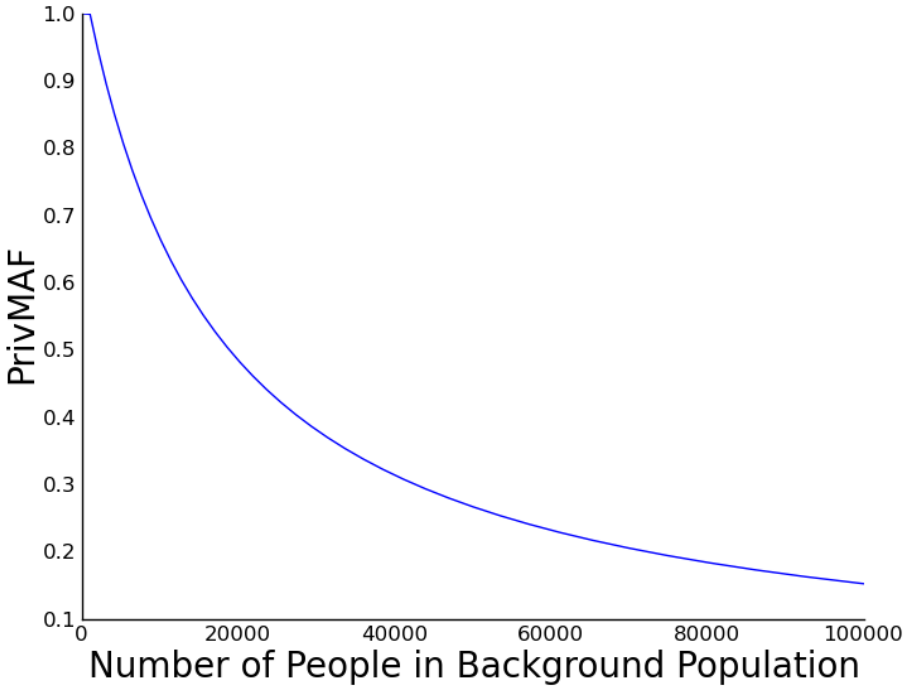

Figure S5: PrivMAF versus background population size. Here we look at how the background population size, denoted by N and graphed on the x axis, affects our privacy measure PrivMAF, which is on the y axis. We see that as the size of the population from which our study is drawn increases the probability of re-identification decreases sharply. This is done with n=1000 study participants and m=200 SNPs.

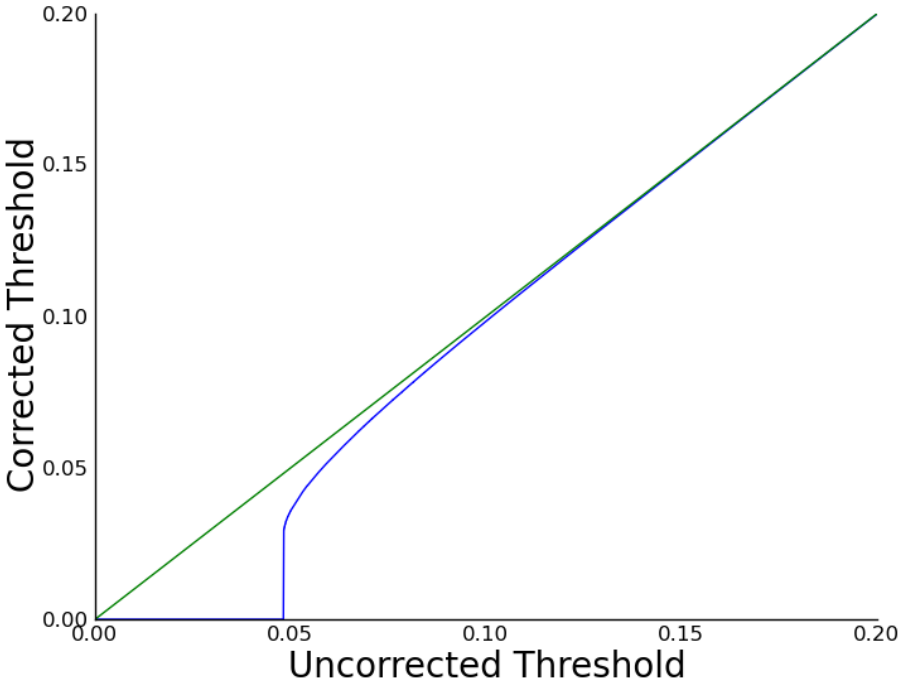

Figure S6: ALGT applied to the WTCCC dataset. A graph of the uncorrected threshold, α, versus the corrected threshold, $\beta = \beta(\alpha)$, from ALGT is given in blue. The green line corresponds to an uncorrected threshold. We see that for some choices of α, correction may be desired. For example, for $\alpha = .05$ the corrected threshold is approximately $\beta = .03$. Here we again use the British Birth Cohort with n=1000 study participants, m=1000 SNPs, and a background population of size N=100,000.