Benchmarking Privatization for GWAS Genomic Data Sharing: Perturbed Allele Frequencies via DP vs. LDP-via-GRR Mechanisms

Eylül Sevinç Sevinç - 83114
Erkmen Erken - 84114

Both group members contributed equally to the implementation of the project. Our implementation and our final comparison results can be found in: https://github.com/eylulssevinc/Privatization-for-GWAS-Genomic-Data-Sharing-COMP-430.git The data downloading can be performed by running download_process_data.ipynb file that can be found under src/download_preprocess.

1. INTRODUCTION AND BACKGROUND

Genomic data has become one of the most valuable data resources in modern science. It supports large scale computational modeling of living systems and allows researchers to relate biological variation to phenotypes (observable traits or outcomes such as disease status or height), advance precision medicine in which medical treatment can be specialized for individuals, and accelerate drug discovery and therapeutic development. In practice, a major scientific tool behind these discoveries is the genome-wide association study (GWAS), where scientists compare groups of people to discover genetic variants that appear statistically associated with that phenotype [1]. Because full genomes are huge and extremely sensitive, GWAS results are commonly shared as summary statistics, especially allele frequencies, rather than raw genotypes, since allele frequencies is a compact, aggregate description of a cohort. Our project is built around this real-world application: *releasing allele-frequency summaries while trying to protect the privacy of the individuals who participated in the study.*

However, human genomic data is not like typical tabular data. A genome is both highly identifying and hard to truly anonymize. Even when explicit identifiers are removed, genetic variants can often be linked back to individuals using auxiliary information. Importantly, unlike a password, a genome cannot be reset after exposure. This makes privacy failures particularly dangerous. A key security concern is that even aggregate releases can leak membership information: an attacker may be able to infer whether a target person is part of the study. This is called a membership inference attack (MIA). Prior work showed that allele frequencies across many SNPs can provide enough signal to test if a specific person contributed to the cohort, yet there is still no widely established, consensus approach for privatizing genomic summary data that reliably balances formal privacy guarantees with GWAS utility [2].

Prior research has shown that genomic privacy risks can persist even when only aggregate GWAS-style statistics are released. For example, Homer et al. showed that allele-frequency information aggregated across many SNPs can be used to test whether a known individual contributed to a mixture/cohort, which would allow the identification or membership-style inference [3]. Subsequent work by Wang et al. further demonstrated that even a relatively limited set of GWAS summary statistics can leak sensitive information (including identity and trait-related inferences) when an attacker can combine the released data with auxiliary data [5].

Within this broader context, on this project, we focus on privatizing case-cohort allele-frequency releases under realistic threat models and clarifying their trade-offs for privacy and utility. For utility, the released statistics should remain accurate enough to preserve downstream scientific conclusions (e.g., which SNPs appear significantly associated with the phenotype). For privacy, the release should make membership inference and similar attacks difficult, ideally under a formal guarantee. Therefore, we decided to employ differential privacy (DP) and local differential privacy (LDP) for this task. They provide a formal privacy protection by introducing controlled randomness. In our project, we did the implementation for three practical ways to privatize GWAS allele-frequency releases and benchmarked them by evaluating how they behave under the two competing goals: reducing membership inference risk while preserving utility of the GWAS data.

1.1. Fundamental Genetics Concepts

DNA and genomic positions: Human DNA is a long sequence over the alphabet {A, C, G, T}, each of these letters representing a nucleotide. A genomic variant is a position where individuals may differ in terms of their DNA. The most common variant type used in GWAS is the single nucleotide polymorphism (SNP): a single-letter change at a specific genomic position that is common enough in the population to study statistically [6].

Alleles and genotypes. At a SNP position, the possible letters (e.g., A or G) are called alleles. Each person has two copies of each chromosome (one from each parent), so at each SNP they have a genotype: two alleles combined (e.g., AA, AG, or GG). If the two alleles are the same, the genotype is *homozygous*; if different, *heterozygous* [7].

In many GWAS computations, genotypes are encoded as a minor-allele count (e.g., assuming the presence of A is rarer than the presence of G, the minor allele is A):

- 0 = no copies of the minor allele (In our example: GG)
- 1 = one copy (In our example: AG)
- 2 = two copies (In our example: AA)

This same 0/1/2 encoding is what our LDP mechanism uses locally for GRR.

Minor allele frequency (MAF): The allele frequency of an allele is how common it is in a given group. The minor allele is the less common allele at that SNP (in a reference population), and MAF is the minor allele's frequency. Allele frequencies are often shared because they summarize a cohort without listing anyone's full genotype, and GWAS papers and downstream analyses often rely on them [7].

GWAS (genome-wide association study) is a case–control association test. A typical GWAS compares a case cohort (people with the phenotype) and a control cohort (people without it) to identify SNPs where allele frequencies differ more than expected by chance. A standard per-SNP approach is a chi-square association test on allele counts or genotype counts, producing p-values and a ranked list of significant SNPs. Our project uses chi-square testing and significant-SNP overlap as a measure of whether privatization preserves GWAS signal [8].

Haplotypes, linkage disequilibrium, and haplotype blocks: Nearby SNPs are not independent. Due to the nature of inheritance, adjacent SNPs are often inherited together. This is called linkage disequilibrium. A haplotype is a common pattern of alleles across a stretch of SNPs; a haplotype block is a region where only a small set of haplotypes is commonly observed. This correlation is an important consideration for our Method 2. Genomic variants within a region are statistically dependent due to linkage disequilibrium, so treating SNPs as independent and perturbing each one separately can disrupt biologically meaningful patterns that downstream analyses rely on, and produce misleading privacy assessments, since an adversary can use the correlation structure to reconstruct unobserved signals from the released noisy statistics [9].

1.2. Differential Privacy (DP) vs Local Differential Privacy (LDP)

Differential privacy (DP) and local differential privacy (LDP) provide two frameworks for limiting information leakage about any single participant. In the central DP setting, a randomized release A(D) is computed from the full dataset D such that for any pair of neighboring datasets D and D' differing in one individual, the probability of any output event changes by at most a factor e^{ε}; $Pr[A(D) \in S] / Pr[A(D') \in S] \leq e^{\varepsilon}$ for all measurable sets S [10]. This bounds the influence of any single record on the released distribution and is why DP is often described as a defense against membership inference, since no adversary should gain substantially more evidence about an individual’s participation from the data than they could without it.

A standard DP tool is the Laplace mechanism, which is adding Laplace noise scaled to the sensitivity of the query (how much one person can change the result). In the GWAS context, releasing allele frequencies is equivalent to answering many count queries (one per SNP) where each query counts how many minor alleles appear in the cohort [11]. Because any single individual can change the count at a given SNP by at most two (they carry 0, 1, or 2 copies), the per-SNP sensitivity is bounded, which allows Laplace noise to be calibrated directly to each SNP’s count.

In contrast, local differential privacy (LDP) does not require trusting a server for privatization. Each participant randomizes their own record locally before sending it to the server. Formally, a local mechanism A is ε-LDP if, for any two possible inputs x and x' and any output y, $Pr[A(D)=y] / Pr[A(D')=y] \leq e^{\varepsilon}$ [12]. As a result, the server receives only privatized reports rather than raw genomic data, which provides protection even if the server is compromised. The trade-off is that local randomization introduces more variance, so accurate estimation typically requires larger sample sizes or accepts reduced utility compared to central DP. In genomics, this comparison is further complicated by linkage disequilibrium, as previously described, is how nearby SNPs are correlated, so treating loci as independent can either lead to loss of biological meaning or allow an adversary to exploit correlation structure. This correlation is an important consideration for our Method 2, where we perturb haplotype-block counts instead of SNP frequencies while maintaining a formal DP guarantee.

1.3. Project Objective and Study Overview

These considerations motivate our project’s goal to quantify how much privacy protection can be gained when releasing GWAS-style allele-frequency summaries, and what utility is lost in the process. In the project, we implement and compare the central DP by adding Laplace noise directly to per-SNP allele counts, central DP that perturbs counts of haplotype blocks, and local DP using generalized randomized response (GRR) where each individual privatizes their response before sending anything to the server so that the server never receives full genotype vectors. We evaluate privacy using a likelihood-ratio style membership test and evaluate utility by how well privatized case frequencies preserve GWAS association signals against a public control group (significant-SNP overlap under chi-square association testing). This allows us to directly compare their behavior under realistic membership-inference risk while measuring how well key GWAS signals are preserved.

2. TECHNICAL APPROACH

Our implementation follows the same overall pipeline stated in the proposal: (i) acquire public genomic data, (ii) construct case/control-style allele-frequency releases for two target regions, (iii) apply three privatization mechanisms (two central DP, one LDP), and (iv) evaluate privacy risk with an LR-style membership test and utility with GWAS-style chi-square signal recovery.

2.1. Implementation Organization and execution flow

The codebase is organized around several notebooks that match the pipeline stages. The first notebook, download_process_data.ipynb, is responsible for data acquisition and preprocessing: downloading the relevant HapMap subsets, extracting the two target regions, converting genotypes into a minor allele count representation, and producing the case and control matrices for downstream mechanisms. The second notebook, method1_laplace.ipynb, implements the baseline central-DP mechanism by adding Laplace noise directly to the per-SNP minor-allele frequency (MAF) vector, rather than perturbing raw allele count histograms. The third notebook, method2_haploblock.ipynb, implements the central-DP mechanism that perturbs haplotype counts within haplotype blocks, then converts the noisy haplotype frequencies back into per-SNP allele frequencies as weighted sums. Finally, method3_grr_ldp.ipynb implements local differential privacy via GRR over genotype values {0,1,2}, allowing each participant to report one or a small fixed number of SNPs and then debiasing the aggregated reports to estimate MAFs.

2.2. Dataset and Cohort Construction from HapMap Genotype Data

We built our experiments on publicly available genotype data from the International HapMap Project (Phase III) hosted on NCBI's HapMap FTP repository [13]. HapMap provides genotypes for multiple labeled population cohorts. In our implementation we treat CEU (Utah residents with Northern and Western European ancestry) as the public control/reference cohort, and we treat MKK (Maasai in Kinyawa, Kenya) as the sensitive, case cohort whose summary statistics are released. Although our proposal initially stated to use Personal Genome Project (PGP) data for the case cohort, we decided on using entirely HapMap data because the PGP chromosome sets were too different from the HapMap chromosome windows we were working with, making consistent SNP matching across cohorts unreliable. The raw HapMap files are organized per chromosome and population and contain many individuals' genotypes at many variant sites. From these files, we extract two fixed genomic windows, one on chromosome 2 (about 5 Mb wide, yielding ~311 SNPs) and one on chromosome 10 (about 1 Mb wide, yielding ~610 SNPs), where a SNP (single-nucleotide polymorphism) means a genomic position at which people commonly differ by a single DNA letter.

For each window and cohort, we convert the raw genotype calls into a simple matrix representation suitable for computation: rows correspond to individuals, columns correspond to SNPs, and each entry is encoded as 0/1/2, which indicates how many copies of the minor allele this person carries at that SNP (0 = none, 1 = one copy, 2 = two copies). From these genotype matrices, we compute per-SNP allele counts and minor allele frequencies (MAFs) (i.e. the ratio of minor alleles in the cohort at each SNP) which is the aggregate release format we aim to privatize.

Since both the privacy attack (LR-based membership scoring) and the utility evaluation (GWAS-style comparisons) assume that "SNP j" refers to the same locus across cohorts, we restrict CEU and MKK to their common SNP identifiers and create an identical SNP ordering before producing any frequency vectors. Finally, we split the MKK individuals into two groups: one subset is used to construct the released case MAF vector (members), and the held-out subset is used as non-members when measuring whether an attacker can tell who participated, while CEU remains as the public reference distribution.

2.3. Preprocessing and cohort construction

The preprocessing notebook creates the processed datasets in three main steps: (1) download and locate the HapMap cohort files for CEU (public control) and the MKK population (case), (2) extract the two target regions and perform genotype encoding, and (3) match the cohorts by keeping only the SNPs they share and putting those SNPs in the exact same order, so that each position refers to the same genetic site in both datasets. SNP alignment is important for both the LR-based privacy metric and the chi-square utility metric, because both assume that ‘SNP j’ refers to the same genomic locus in case and control. After alignment, we construct the membership inference setting by splitting the case cohort (MKK) into two disjoint subsets: a member set used to compute the released (privatized) case allele frequencies, and a non-member set held out as individuals not included in the release for the LR test, while CEU remains the public reference/control cohort used as the baseline frequency distribution. For method 2, preprocessing also ensures the data can be expressed in haplotypes, so we can build haplotype-block counts and then convert the noisy haplotype frequencies back into SNP-level allele frequencies.

2.4. Method 1: Naïve SNP-level Laplace perturbation (central DP baseline)

Method 1 serves as a global-DP baseline by adding independent Laplace noise to each SNP’s released statistic. Rather than adding noise to the full vector of raw allele counts with a “2N” histogram-style sensitivity, our implementation perturbs the per-SNP minor-allele frequency (MAF) vector directly, which is the actual release format used throughout the project. We first compute the case-cohort MAFs from the member split of the case population (MKK), then add independent Laplace noise to each SNP’s frequency and clip the resulting values to the valid range [0,1]. Because downstream GWAS-style utility evaluation requires non-negative inputs, the implementation also applies basic safeguards when converting the noisy frequencies into the count quantities needed for chi-square computations. Method 1 is a stable baseline for understanding how much privacy protection can be obtained from naive SNP-perturbation of the released statistics.

2.5. Method 2: Haplotype-block Laplace perturbation (central DP)

As discussed previously, SNPs within a genomic region are correlated due to linkage disequilibrium, so perturbing each SNP independently (Method 1) can distort biologically meaningful joint patterns and can also give an adversary structure to exploit when interpreting the released noisy data. Method 2 therefore perturbs a correlation-preserving representation based on haplotype blocks: instead of adding noise at each SNP, we operate on the distribution of haplotypes (common allele patterns) within each block. In the project, for each block b, we construct a histogram of case haplotype counts from the member split of the case cohort (MKK), add Laplace noise with a block-specific budget ε_b, normalize the noisy histogram to obtain noisy haplotype frequencies, and then derive each SNP’s perturbed minor allele frequency as a weighted sum of those noisy haplotype frequencies (weights determined by whether each haplotype carries the minor allele at that SNP). We allocate the total privacy budget ε across blocks so that $\sum_b \varepsilon_b = \varepsilon$, using a rule proportional to the number of haplotypes represented per block. Two practical implementation details are important for robustness. First, haplotype histograms can become huge and extremely sparse, so for each block we keep only the Top-K most common haplotypes and group everything else into an “OTHER” bucket. When we convert back to SNP-level allele frequencies, we account for the OTHER bucket by assigning its remaining probability mass using the CEU control MAF as a reasonable baseline, so the final frequency vector stays well-defined. Second, haplotype-block mappings do not always line up perfectly with the final SNP subset used for a region; when a block cannot be mapped cleanly, we fall back to the corresponding control-based frequency for that block to avoid producing invalid outputs. Finally, because filtering can leave different numbers of usable MKK individuals in the chromosome 2 and chromosome 10 regions, we perform the member/non-member split separately for each region rather than forcing a tiny intersection across both chromosomes.

2.6. Method 3: LDP via generalized randomized response (GRR) with debiasing

Method 3 implements a local differential privacy (LDP) mechanism based on generalized randomized response (GRR), where privatization occurs on the participant side rather than at a trusted curator. Instead of transmitting a full genotype vector, each individual reports only a small amount of information: they select a SNP index at random and encode their genotype at that locus as a three-valued minor-allele count in {0,1,2}. In our implementation, each participant can report either one SNP or a small fixed number k of SNPs to reduce noise in the final estimates. The participant then applies GRR locally: with probability determined by the privacy budget ε, they report the true value, and otherwise they report one of the other two values according to the GRR rule. The server receives only the SNP indices and the randomized categorical response(s), aggregates reports SNP-by-SNP, and then applies the standard GRR debiasing step to recover an unbiased estimate of the genotype distribution at each SNP, which is finally converted into an allele-frequency estimate. The server never sees raw genotypes, so this approach does not rely on trusting the curator. The cost in method 3 is that the local randomization adds noise per participant, so the frequency estimates are typically noisier than central DP for the same ε, unless the cohort is large.

2.7. Evaluation pipeline:

We measure privacy risk using a likelihood-ratio (LR) membership test that combines information across many SNPs into a single LR score for each individual. In our setup, the released case frequencies (either unperturbed or privatized) define the “case” model $\hat{p}_j$, while the public CEU cohort provides the reference frequencies p_j. We then test membership within the same case population by comparing individuals who were included in the released summary (members, the case train split) against held-out individuals from that same population (non-members, the case test split). We choose the decision threshold as the 95th percentile of the non-member LR scores (equivalently a 5% false positive rate), and report LR power as the fraction of members whose LR score exceeds this threshold.

$$L(i) = \sum_{j=1}^{m} \left[\log P(g_{ij} \mid \hat{p}_j) - \log P(g_{ij} \mid p_j) \right]$$

Figure 1: Genotype-based log-likelihood ratio (LR) membership score. For an individual i, the score L(i) aggregates evidence across m SNPs by comparing how likely the individual’s observed genotypes $\{g_{ij}\}$ are under the released case-cohort allele frequencies $\{\hat{p}_j\}$ versus the public reference/control allele frequencies $\{p_j\}$. Here $\{g_{ij}\}$ in {0,1,2} denotes the minor-allele count of individual i at SNP j. Assuming Hardy–Weinberg equilibrium, genotype probabilities are:

$$P(g=0 \mid p) = (1-p)^2 \text{ , } P(g=1 \mid p)\ 2p(1-p),\ P(g=2 \mid p) = p^2$$

Larger values of L(i) indicate that the individual's genotypes are more consistent with the case cohort than with the reference cohort, and are therefore interpreted as stronger evidence of membership.

For utility, we use GWAS-style signal recovery: we compute per-SNP chi-square association statistics from allele-count tables (minor vs. major allele) comparing the case cohort to the public control cohort, call SNPs significant using a fixed p-value cutoff (10^{-3}), and evaluate how closely the significant set from privatized case frequencies matches the significant set from unperturbed case frequencies using overlap metrics (Jaccard). To keep these downstream computations stable under noise, we bound privatized frequencies to range [0,1] and reconstruct allele counts for the chi-square test.

RESULTS AND DISCUSSION

To evaluate the three privatization mechanisms consistently across methods and also make it easy to compare, we measured the privacy risk, data distortion, and GWAS-level utility on two fixed genomic regions (chr2_5Mb and chr10_1Mb). For each region, we treated the case cohort as the sensitive dataset and computed its true per-SNP minor allele frequencies (MAFs), and each method then released a privatized version of the case MAF vector under a privacy budget ε. We assessed the privacy using an LR based membership inference test, where we computed an LR score for each individual and defined the LR power as the fraction of member individuals (case-train) whose LR score exceeds the 95th percentile of nonmember individuals (case-test). In hypothesis, under stronger privacy (more noise and smaller ε), we expected LR power to stay close to the false positive rate of 0.05. However, with larger ε, we expected this to make the attack more effective and increase the LR power. We measured the utility by how well the privatized releases preserved the downstream GWAS conclusions against the public control cohort, using GWAS overlap (Jaccard) between significant SNP sets from the true vs privatized case-control comparison. Finally, to quantify direct perturbation strength independent of GWAS thresholding, we also measured the distortion as MAF MAE (mean absolute error between true and privatized case MAFs). In an ideal privacy utility tradeoff, with increasing ε, we would expect decreasing distortion (lower MAE), generally increased GWAS overlap, and as privacy gets weaker with higher ε, an increased LR power. In addition to our hypothetical expectations, it should also be taken into account that the finite sample size, region specific structure, and randomness in the data could also introduce nonmonotonicity and different behavior between chr2 and chr10.

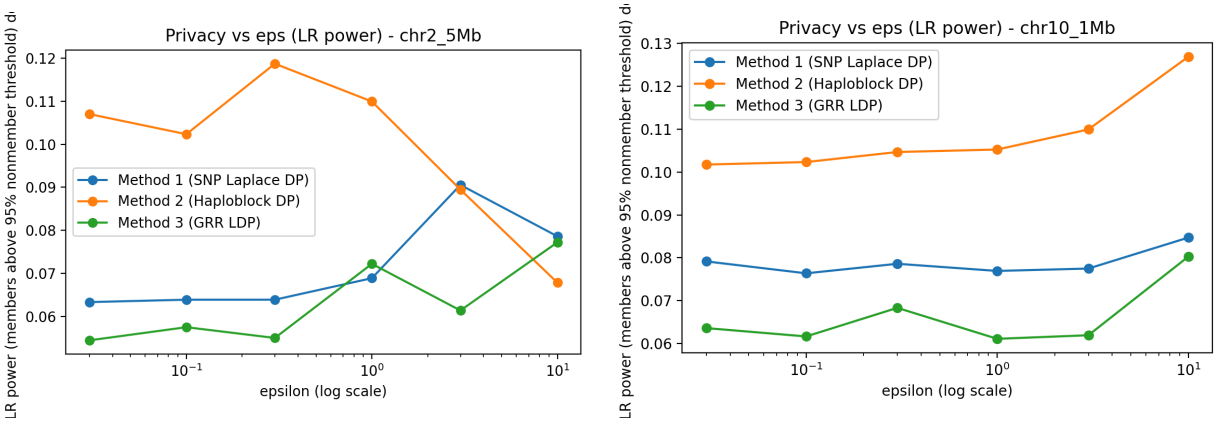

Figure 2. Privacy vs ε under LR based membership inference (LR Power). Left: chr2_5Mb. Right: chr10_1Mb. Each curve shows LR power (fraction of member individuals whose LR score exceeds the 95th percentile nonmember threshold) as a function of the privacy budget ε (log scale). Lower values indicate stronger privacy, and harder membership inference attacks. Curves compare Method 1 (SNP level Laplace DP on released MAFs), Method 2 (haplotype-block Laplace DP), and Method 3 (GRR-based LDP).

Table 1. Quantitative Summary of privacy results presented in Figure 2.

Method	chr2 ε=0.1	chr2 ε=1.0	chr2 ε=10	chr10 ε=0.1	chr10 ε=1.0	chr10 ε=10
Method 1 (SNP Laplace DP)	0.06389 ± 0.037	0.06889 ± 0.047	0.07861 ± 0.052	0.07639 ± 0.053	0.07694 ± 0.047	0.08472 ± 0.040
Method 2 (Haploblock DP)	0.1023 ± 0.060	0.1099 ± 0.066	0.06784 ± 0.026	0.1023 ± 0.061	0.1053 ± 0.063	0.1269 ± 0.058
Method 3 (GRR LDP)	0.0575 ± 0.034	0.07222 ± 0.048	0.07722 ± 0.041	0.06167 ± 0.044	0.06111 ± 0.026	0.08028 ± 0.041

The two privacy plots presented in Figure 2 summarizes how vulnerable each released case-allele-frequency is to our LR-based membership inference test as the privacy budget ε increases. In our evaluation, LR power measures the fraction of true members whose LR scored exceed a decision threshold set at the 95th percentile of nonmembers, therefore, lower LR power indicates stronger privacy, and the attack cannot reliably distinguish members from nonmembers. Conceptually we would expect that with larger ε, the membership inference should get easier and thus the LR power should increase. Although, it should also be taken into account that our setting has finite samples and involves randomness, so perfectly monotone behavior is not always guaranteed.

Across both chromosomes, Method 3 (GRR LDP) shows the most consistent privacy, with consistently lower LR scores than other methods. This method achieved the lowest LR power at the smaller ε values, which was also expected. Method 1 (SNP Laplace on the released MAF vector) is also reasonably stable, as ε increases, the LR power increases mildly for both chromosomes. As apparent in Figure 2, method 2 (Haploblock DP) yielded the most mixed and most chromosome-dependent results. On chr10, method 2 behaved according to our expectations, and as ε increased, the LR power also increased. On chr2, however, method 2 is non-monotonic in ε, which suggests that this region is more sensitive to small cohort effects for the haplotype block pipeline we used. This result shows that with our current dataset size for the chr2 region, the membership inference outcome becomes less predictable.

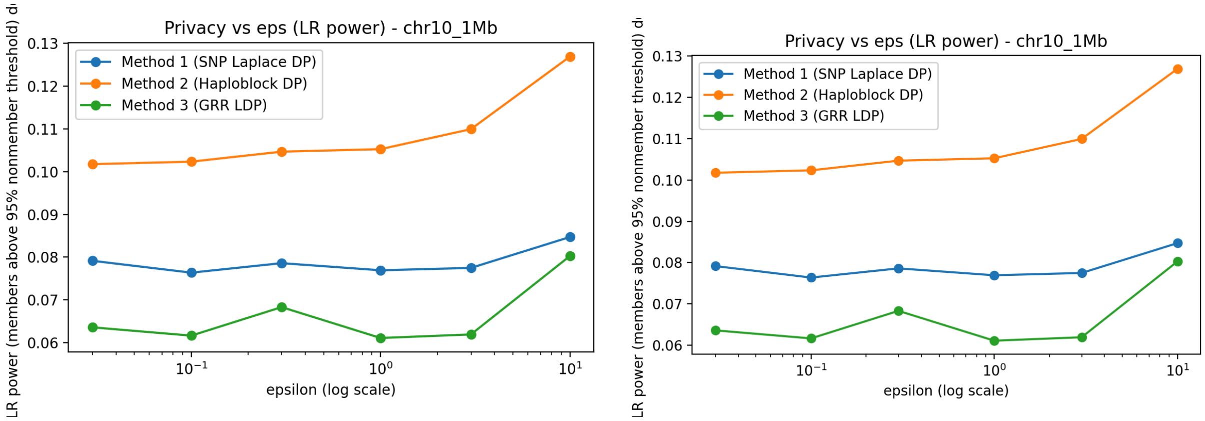

Figure 3. Utility vs ε measured by GWAS overlap (Jaccard). Left: chr2_5Mb. Right: chr10_1Mb. Each curve shows GWAS overlap utility as Jaccard(sig_true, sig_noisy), where sig_true is the set of significant SNPs from the true case-control GWAS and sig_noisy is the set from the GWAS computed using the privatized case release against the same public control. The x axis is the privacy budget ε (log scale). Higher values indicate better utility (greater agreement with the true GWAS signal). Curves compare Method 1 (SNP level Laplace DP on released MAFs), Method 2 (haplotype-block Laplace DP), and Method 3 (GRR-based LDP).

Table 2. Quantitative summary of utility results presented in Figure 3.

Method	chr2 ε=0.1	chr2 ε=1.0	chr2 ε=10	chr10 ε=0.1	chr10 ε=1.0	chr10 ε=10
Method 1 (SNP Laplace DP)	0.5447 ± 0.022	0.5414 ± 0.021	0.5395 ± 0.023	0.5108 ± 0.016	0.5059 ± 0.017	0.4842 ± 0.020
Method 2 (Haploblock DP)	0.1586 ± 0.077	0.1094 ± 0.049	0.1225 ± 0.011	0.2267 ± 0.045	0.1937 ± 0.054	0.1599 ± 0.024
Method 3 (GRR LDP)	0.2469 ± 0.023	0.2350 ± 0.023	0.2152 ± 0.025	0.1386 ± 0.016	0.1393 ± 0.019	0.1293 ± 0.018

Figure 3 and Table 2 presents how well each privatized release preserved downstream GWAS conclusions as the privacy budget ε changes. As previously explained, we measured the utility as the GWAS overlap (Jaccard) between the set of significant SNPs obtained from the true case-control comparison and the set obtained when the privatized case-group MAFs are used against the same public control cohort. Therefore, higher Jaccar indicates better utility, meaning a more similar set of significant SNPs compared to the unperturbed analysis. Hypothetically, we would expect that larger ε improves utility. It is also important to note that the Jaccard overlap is a threshold based metric, and if noise changes the number of SNPs that are declared significant, the overlap can shift in a way that is not perfectly smooth or monotone under finite sample sizes like ours.

As it can be seen from Figure 3 and Table 2, across both chromosomes, Method 1 (SNP Laplace DP) preserved utility the best by a clear margin, and its GWAS overlap remained high and stable. So, as expected, adding noise directly at the SNP level to the released MAF preserved the overall per-SNP signal structure. However, in Method 2 and Method 3, lower overlap values were achieved, which means a stronger disruption of which SNPs cross the significance threshold. Method 2 typically performed better than Method 3 on both chromosomes. In contrast to the expectation of achieving better utility with bigger ε values, both Method 2 and 3 showed fluctuations. This could also mean that the “significant SNP set” approach is sensitive in the context of this project. As ε changes the privatized analysis may produce a different number of significant SNPs and thus the Jaccard can drop. In practice, this problem can be stabilized by using data with bigger cohort sizes, as this would reduce the threshold sensitivity.

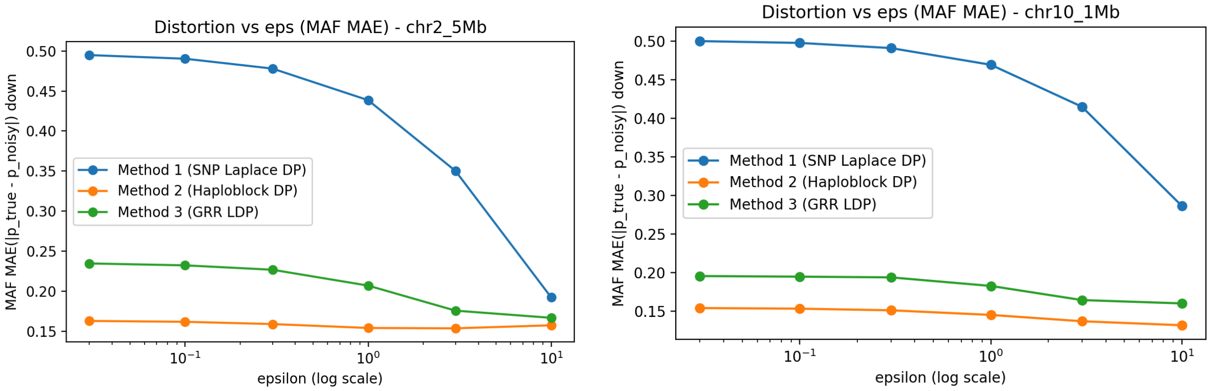

Figure 4. Distortion vs ε (MAF MAE) for the released case allele frequencies. Left: chr2_5Mb. Right: chr10_1Mb. Each curve shows the mean absolute error (MAE) between the true case minor allele frequency (MAF) vector and the privatized case MAFs as a function of the privacy budget ε (log scale). Higher MAE indicates a more perturbed release. Curves compare Method 1 (SNP level Laplace DP on released MAFs), Method 2 (haplotype-block Laplace DP), and Method 3 (GRR-based LDP).

Figure 4 summarizes how much each method changes the released case allele frequency vector. For each ε, the mean absolute error (MAE) is computed between the true case minor allele frequencies (MAFs) and the privatized MAFs. In the plots, larger MAE means that the release is more distorted with the applied method of privatization, and it measures the direct effect of the noise added. As ε increases (weaker privacy and less noise), the expectation is that MAE should be decreasing, since the privatized frequencies should become closer to the true frequencies.

Across both chr2_5Mb and chr10_1Mb, Method 1 showed the strongest ε dependency. When ε was small, the SNP level Laplace mechanism added relatively large noise to many SNPs, so the released MAF was heavily distorted, and the MAE was calculated to be high. As ε increased, the noise scale decreased, which meant the the privatized MAFs got similar to the true case MAFs.

In method 3, the overall trend can also be observed, but the change is more gradual. Since GRR is a local mechanism, noise was introduced at the individual level and the released case frequencies were obtained through an estimation step. As ε increased, each individual’s result got less random, so the estimated MAF vector became closer to the true case MAFs. Compared to method 1, however, the MAE curve for method 3 did not decline as sharply, since the frequency estimation step still carries finite sample variability. Method 2 behaved differently from both method 1 and 2, its MAE remained comparatively low and flat across ε on both chromosomes. This result occurred because of the technical logic of method 2, which was adding laplace noise to haplotype block counts and then reconstructing the SNP level MAFs as weighted sums over noisy frequencies. The block to SNP reconstruction step that we had likely acted like a smoothing step, and it kept the average per-SNP frequency deviation small even when the distribution has been perturbed.

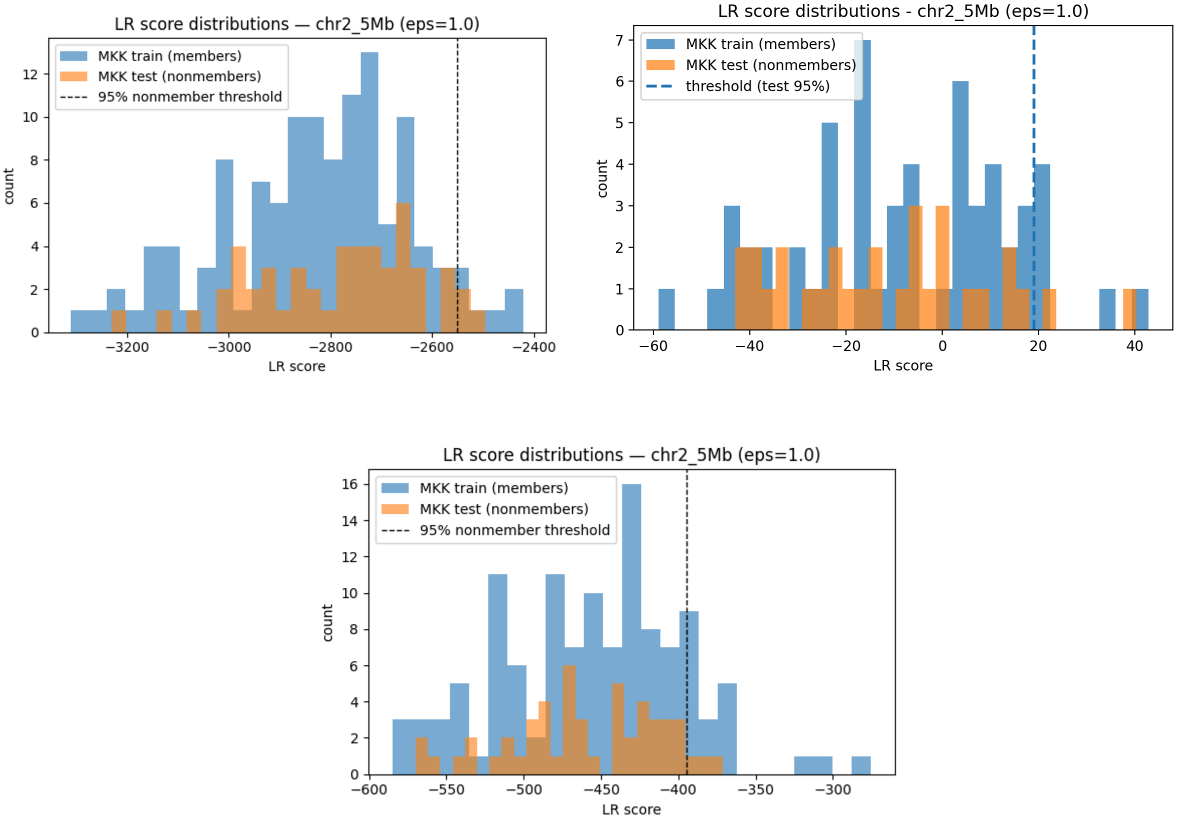

Figure 5. LR-score distribution histograms for chr2 at ε =1.0 (membership inference snapshot). Top-left: Method 1 (SNP level Laplace DP on released MAFs). Top-right: Method 2 (haplotype-block Laplace DP). Bottom: Method 3 (GRR-based LDP). In each panel, blue bars show LR scores for members (MKK train) and orange bars show LR scores for nonmembers (MKK test). The dashed vertical line indicates the 95th percentile nonmember threshold used to compute LR power; members to the right of this line are classified as members after the membership inference attack.

In Figure 5, the LR score distribution histograms can be seen at ε = 1.0 on the chr2 region. The histograms provide an intuitive view of how each privatization method affects the membership inference. In each plot, the blue distribution shows the LR scores for members (MKK train) and the orange distribution shows nonmembers (MKK test), while the dashed vertical line shows the 95th percentile nonmember threshold used by our LR power metric. In an ideal privacy-preserving release, the member and nonmember score distributions should be overlapping well and only a small number of members should be to the right of this threshold line. Across all three methods, we observed a noticeable overlap between members and nonmembers in all cases, but with different overlapping extent and shape. Method 1 showed a broad member distribution with some of them extending toward the threshold, which also shows that membership inference was not eliminated, but it was partially constrained. Method 2 yielded a wide overlap between the two groups, which was aligning with the fact that using noisy haplotype frequencies made it harder to distinguish both members and nonmembers. Method 3 similarly produced overlapping distributions, and it appeared to be the most privacy preserving method overall, consistent with the privacy comparison in Figure 2.

To conclude, it can be said that to achieve a more meaningful privatization in the GWAS genomic data sharing context, the right statistical structure (biological utility) should be preserved while also privatizing the data to reduce the re-identification risk. Both of these objectives can be more reliable when the data (cohort and region presented in the GWAS data) provides stronger and less noisy signals. With larger within-population cohort sizes and ideally, larger genomic windows, the allele frequency estimations and test statistics would make the membership inference evaluation better by avoiding the randomness. Overall, our benchmark shows that to meaningfully privatize GWAS data, it is required to have both a principled mechanism and enough data support for the preserved signal to remain robust under noise.

REFERENCES

[1] Lin, Z., Owen, A. B., & Altman, R. B. (2004). Genomic research and human subject privacy. *Science*, *305*(5681), 183-183.

[2] Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J. P., ... & Wang, X. (2015). Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, *48*(1), 1-44.

[3] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, *4*(8), e1000167. https://doi.org/10.1371/journal.pgen.1000167

[4] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. 2009. Learning your identity and disease from research papers: information leaks in genome wide association study. In Proceedings of the 16th ACM conference on Computer and communications security (CCS '09). Association for Computing Machinery, New York, NY, USA, 534–544. https://doi.org/10.1145/1653662.1653726

[5] Uhlerop, C., Slavković, A., & Fienberg, S. E. (2013). Privacy-Preserving Data Sharing for Genome-Wide Association Studies. *The Journal of privacy and confidentiality*, *5*(1), 137–166.

[6] Popovic A, Orrick JA. Biochemistry, Mutation. [Updated 2024 Sep 10]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK576397

[7] National Human Genome Research Institute. (n.d.). *Allele*. Genetics Glossary. Retrieved January 21, 2026, from https://www.genome.gov/genetics-glossary/Allele

[8] Kai Wang, Statistical tests of genetic association for case–control study designs, *Biostatistics*, Volume 13, Issue 4, September 2012, Pages 724–733, https://doi.org/10.1093/biostatistics/kxs002

[9] Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science (New York, N.Y.)*, *296*(5576), 2225–2229. https://doi.org/10.1126/science.1069424

[10] Central DP definition (neighboring datasets; eεeε bound over all events SS)
Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis.* In Theory of Cryptography Conference (TCC 2006).

[11] Uhlerop, C., Slavković, A., & Fienberg, S. E. (2013). Privacy-Preserving Data Sharing for Genome-Wide Association Studies. *The Journal of privacy and confidentiality*, *5*(1), 137–166.

[12] Local Differential Privacy definition (randomization at the user; εε-LDP ratio bound for all x,x′,yx,x′,y)
Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2011). *What Can We Learn Privately?*SIAM Journal on Computing.

[13] International HapMap Consortium. (n.d.). *HapMap genotypes* [Data set]. National Center for Biotechnology Information, National Library of Medicine. https://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/

[14] Visscher, P. M., & Hill, W. G. (2009). The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS genetics*, *5*(10), e1000628. https://doi.org/10.1371/journal.pgen.1000628

[15] Johnson, A., & Shmatikov, V. (2013, August). Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1079-1087).

[16] Jiang, X., Zhao, Y., Wang, X., Malin, B., Wang, S., Ohno-Machado, L., & Tang, H. (2014). A community assessment of privacy preserving techniques for human genomes. *BMC medical informatics and decision making*, *14*(Suppl 1), S1.

[17] National Center for Biotechnology Information. (n.d.). *HapMap data (FTP directory)* [Data set]. Retrieved December 25, 2025, from https://ftp.ncbi.nlm.nih.gov/hapmap/

[18] Sankararaman, S., Obozinski, G., Jordan, M. I., & Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature genetics*, *41*(9), 965–967. https://doi.org/10.1038/ng.436

[19]Bonomi, L., Huang, Y., & Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nature genetics*, *52*(7), 646-654.

[20] Chen, J., Wang, W. H., & Shi, X. (2020). Differential privacy protection against membership inference attack on machine learning for genomic data. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium* (pp. 26-37).