# SAFETY: Secure gwAs in Federated Environment Through a hYbrid solution
# Appendix

## A Statistical Tests Used in GWAS

In this section, we discuss the required background for genomic data and the statistical tests that we used for GWAS. These statistical tests were also discussed in some of the prior works on privacy-preserving GWAS [1, 2]. We are describing these tests here to keep the article self-contained.

Table 1: Data representation in each party

| | | Sequence | | | | |
|---|---|---|---|---|---|---|
| # | **Case** | **rs4426** | **rs4305** | **rs4630** | | **Cancer** |
| Data Owner 1 | 1 | CC | CT | GG | ... | Negative |
| | 2 | CT | CT | AG | ... | Negative |
| | 3 | CC | CT | GG | ... | Negative |
| Data Owner 2 | 1 | CC | CT | GG | ... | Negative |
| | 2 | CT | CC | GG | ... | Positive |
| | 3 | CC | CT | GG | ... | Positive |
| Data Owner 3 | 1 | CT | CC | AG | ... | Positive |
| | 2 | CT | CT | AG | ... | Negative |
| | 3 | TT | CC | GG | ... | Positive |
| Data Owner 4 | 1 | TT | CC | AA | ... | Positive |
| | 2 | CC | CC | GG | ... | Positive |
| | 3 | CC | CT | GG | ... | Positive |

### A.1 Genomics

Genes may have different forms which are located at a particular position or locus, on a chromosome. The different variants of a gene at a locus are called alleles. This kind of genetic variations in DNA sequences has a significant influence on disease and phenotype. The most common form of genetic variants is Single Nucleotide Polymorphism (SNP). A SNP is a genetic variation, which refers to an alteration of a single nucleotide (A, T, C, or G). A SNP is generally bi-allelic: there are two alleles at a SNP locus.
In humans, chromosomes are inherited in pairs, one from each parent. Hence, at a particular locus, there is an allele in each of the two chromosomes. These two alleles together are called genotype. On the other hand, the sequence of alleles along a chromosome is called a haplotype.

### A.2 Linkage Disequilibrium

Linkage disequilibrium refers to a nonrandom association of alleles at two or more loci. On the other hand, when the alleles are randomly associated they are said to be in a state of linkage equilibrium.

For example, suppose at SNP locus 1, the possible alleles are *A*, *a*, and at SNP locus 2, the possible alleles are *B*, *b*. So, there are four possible pairs of alleles *AB, Ab, aB, ab* which are called haplotypes. Let, $N_{AB}, N_{Ab}, N_{aB}$, and $N_{ab}$ be the number of four haplotypes $AB, Ab, aB$, and $ab$ respectively. If the total size of the population sample is $N$, then the population frequencies are calculated as:
$P_{AB} = \frac{N_{AB}}{N}$, $P_{Ab} = \frac{N_{Ab}}{N}$, $P_{aB} = \frac{N_{aB}}{N}$, $P_{ab} = \frac{N_{ab}}{N}$. Let,

$$P_A = \text{frequency of allele A at locus 1},$$

$$P_a = \text{frequency of allele a at locus 1},$$

$$P_B = \text{frequency of allele B at locus 2,}$$

$$P_b = \text{frequency of allele b at locus 2}$$

If locus 1 and 2 are independent, we would expect to see that a haplotype, say, AB has frequency $P_A P_B$. If the frequency of haplotype AB is different from $P_A P_B$, the two loci are said to be in linkage disequilibrium. Linkage disequilibrium is defined based on following quantity:

$$D_{AB} = P_{AB} - P_A P_B \tag{1}$$

Here, $D_{AB}$ is called the coefficient of linkage disequilibrium. It characterizes the extent to which two alleles are nonrandomly associated. Each pair of alleles has a specific value of $D$. The values of $D$ for different pairs of alleles are constrained by the following facts,

$$P_A + P_a = 1$$

$$P_B + P_b = 1$$

$$P_{AB} + P_{Ab} + P_{aB} + P_{ab} = 1$$

The above set of equations demonstrate the underlying relationship among allele frequencies and haplotype frequencies. This relationship is illustrated in Table 2.

Table 2: Relationship among allele frequencies and haplotype frequencies

| SNP 2 | | | | Total |
|---|---|---|---|---|
| SNP 1 | Allele | B | b | |
| | A | $P_{AB}$ | $P_{Ab}$ | $P_A$ |
| | a | $P_{aB}$ | $P_{ab}$ | $P_a$ |
| Total | | $P_B$ | $P_b$ | 1 |

Only one value of $D$ is required to characterize linkage disequilibrium between two loci because $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$. Generally, $D$ is used without subscript.
The following table illustrates the value of haplotype frequency under linkage equilibrium and linkage disequilibrium.

| $P_{ij}$ | Linkage disequilibrium | Linkage equilibrium |
|---|---|---|
| $P_{AB}$ | $P_{AB} = P_A P_B + D$ | $P_{AB} = P_A P_B$ |
| $P_{Ab}$ | $P_{Ab} = P_A P_b - D$ | $P_{Ab} = P_A P_b$ |
| $P_{aB}$ | $P_{aB} = P_a P_B - D$ | $P_{aB} = P_a P_B$ |
| $P_{ab}$ | $P_{ab} = P_a P_b + D$ | $P_{ab} = P_a P_b$ |

It can be shown that,

$$D = P_{AB} P_{ab} - P_{Ab} P_{aB} \tag{2}$$

The range of $D$ depends on the allele frequencies, which makes it troublesome to use it as a measure of disequilibrium. There are two scaled-down variants: $D'$- measure or $r^2$- measure.

**$D'$-measure**: The $D'$-measure is given by: $D' = |\frac{D}{D_{max}}|$; where $D_{max} = \begin{cases} \min\{P_A P_b, P_a P_B\}, & \text{if } D > 0 \\ \min\{P_A P_B, P_a P_b\}, & \text{if } D < 0 \end{cases}$

**$r^2$-measure**: The $r^2$-measure is given by: $r^2 = \frac{D^2}{P_A P_B P_a P_b}$

The range of both $D'$ and $r^2$ is $[0, 1]$. Here, 0 indicates complete linkage equilibrium and 1 indicates complete linkage disequilibrium.

Now, let us consider the data from Table 1. We try to determine if *rs4305* and *rs4630* are at linkage disequilibrium. Both SNPs are bi-allelic. So, there are four possible haplotypes: CA, CG, TA and TG.

| rs4630 | | | |
|---|---|---|---|
| rs4305 | Allele | A | G |
| | C | CA | CG |
| | T | TA | TG |

Let, total number of haplotypes be $N$ and number of haplotypes CA, CG, TA, and TG are $N_{CA},\ N_{CG},\ N_{TA},$ and $N_{TG}$ respectively. We use subscript $i$ to denote the contribution of $i^{th}$ data owner. So, for $n$ data owners,

$$N_{CA} = N_{CA_1} + N_{CA_2} + \ldots + N_{CA_n}$$

$N_{CG},\ N_{TA}$ and $N_{TG}$ are computed similarly. Now, we can calculate the frequencies of these haplotypes. For instance,

$$P_{CA} = \frac{N_{CA}}{N}$$

With these haplotype frequencies we can compute frequency of a particular allele at a specific locus.

$$\begin{aligned} P_C &= P_{CA} + P_{CG} \\ P_T &= P_{TA} + P_{TG} \\ P_A &= P_{CA} + P_{TA} \\ P_G &= P_{CG} + P_{TG} \end{aligned}$$

So,

$$D = P_{CA}P_{TG} - P_{CG}P_{TA} \tag{3}$$

## A.3 Hardy-Weinberg Equilibrium

Hardy-Weinberg Equilibrium (HWE) is a principle which states that allele and genotype frequencies in a population will remain unchanged from one generation to the next generation given that there are no evolutionary influences. A lot of assumptions are required for HWE to hold including no inbreeding, random mating, infinite population size, and no mutation, selection, or migration [3].

HWE indicates a relationship between allele frequency in parents and genotype frequency in offspring. For example, consider two alleles $A$ and $a$ of locus 1 (mentioned in A.2). After one round of random mating, the genotype frequencies in the offspring are:

$$\left.\begin{aligned} P(\text{genotype } AA) &= P_A^2 \\ P(\text{genotype } Aa) &= 2P_AP_a \\ P(\text{genotype } aa) &= P_a^2 \end{aligned}\right\} \tag{4}$$

If all the genotypes of a population satisfy equation 4, it is said to be in HWE . It is noteworthy that the sum of all the frequencies is the binomial expansion of the square of the sum of $P_A$ and $P_a$. Since $P_A + P_a = 1$, $P_A^2 + 2P_AP_a + P_a^2 = (P_A + P_a)^2 = 1$.

Table 3: A Punnett square demonstrating the probabilities of generating all possible genotypes at locus 1

| | | Females | |
|---|---|---|---|
| Males | Allele | A($P_A$) | a($P_a$) |
| | A($P_A$) | AA($P_A^2$) | Aa($P_AP_a$) |
| | a($P_a$) | Aa($P_AP_a$) | aa($P_a^2$) |

To estimate deviation from HWE, Pearson goodness of fit test is generally used. In this test, the observed genotype counts are obtained from data, and the expected genotype counts are calculated using HWE.

Table 4: Pearson Goodness of Fit Test for HWE

| Genotype | AA | Aa | aa | total |
|---|---|---|---|---|
| Observed count | $n_{AA}$ | $n_{Aa}$ | $n_{aa}$ | $n$ |
| Expected count | $nP_A^2$ | $nP_AP_a$ | $nP_a^2$ | $n$ |

$P_A$ can be calculated using, $P_A = \frac{n_{AA}}{n} + \frac{1}{2} \times \frac{n_{Aa}}{n}$.
Pearson Goodness of Fit Test for HWE is given by:

$$\chi^2 = \sum \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected Count}} \tag{5}$$

Now, let us consider, the data from Table 1. We try to determine if HWE holds for *rs4305*. Possible genotypes

at *rs4305* are CC, CT and TT.
Observed count for genotype CC is given by,

$$n_{CC} = n_{CC_1} + n_{CC_2} + n_{CC_3} + n_{CC_4}$$

$n_{CT}$ and $n_{TT}$ are calculated similarly. So, here population size is,

$$n = n_{CC} + n_{CT} + n_{TT}$$

Now, the frequency of the allele C in this population is calculated as follows:

$$P_C = \frac{n_{CC}}{n} + \frac{1}{2} \times \frac{n_{CT}}{n} \tag{6}$$

Therefore, the frequency of the allele T, $P_T = 1 - P_C$.
Expected counts of genotype CC, CT and TT are $nP_C^2, 2nP_CP_T$, and $nP_T^2$ respectively. Pearson goodness of fit test for HWE is given by:

$$\chi^2 = \frac{(n_{CC} - nP_C^2)^2}{nP_C^2} + \frac{(n_{CT} - 2nP_CP_T)^2}{2nP_CP_T} + \frac{(n_{TT} - nP_T^2)^2}{nP_T^2}$$

In this case, critical chi-square value with 1 degree of freedom (3 genotypes - 2 alleles) is 3.841 (for 0.05 significance level). If $\chi^2 < 3.841$, then HWE holds in this population.

## A.4 Cochran-Armitage Test for Trend

Cochran-Armitage Test for Trend (CATT) is highly used in case-control studies in order to determine if an allele is associated with a disease. Such a study identifies factors that may contribute to a disease by comparing the genotypes of the individuals who have the disease (cases) with the individuals who do not have the disease (controls).
CATT can be applied when the data takes a form of $2 \times k$ contingency table. For instance, a $2 \times 3$ contingency table can be constructed with 3 genotypes vs. cases/controls as follows.

| | AA | Aa | aa | Sum |
|---|---|---|---|---|
| Controls | $N_{11}$ | $N_{12}$ | $N_{13}$ | $R_1$ |
| Cases | $N_{21}$ | $N_{22}$ | $N_{23}$ | $R_2$ |
| Sum | $C_1$ | $C_2$ | $C_3$ | $N$ |

In this table, $N_{ij}$ represents genotype frequency, $R_i$ is the sum of the $i^{th}$ row and $C_j$ is the sum of the $j^{th}$ column. If a contingency table like this is given, CATT computes the trend test statistic as follows:

$$T = \sum_{i=1}^{3} w_i(N_{1i}R_2 - N_{2i}R_1) \tag{7}$$

where $w_i$ are weights. Chi-square value is given by:

$$\chi^2 = \frac{T^2}{Var(T)} \tag{8}$$

where the variance of $T$ is given by:

$$Var(T) = \frac{R_0R_1}{N}(\sum_{i=1}^{3} w_i^2C_i(N - C_i) - 2\sum_{i=1}^{2}\sum_{j=i+1}^{3} w_iw_jC_iC_j) \tag{9}$$

## A.5 Fisher's Exact Test

Like CATT, Fisher's Exact Test (FET) is another statistical test which is used to analyze contingency table in order to find association. FET performs well when the sample size is small.

FET operates on a contingency table to identify any association between a variable of $s$ different categories and a variable with $t$ different categories.

| | Group 1 | Group 2 | $\cdots$ | Group s | Sum |
|---|---|---|---|---|---|
| Category 1 | $N_{11}$ | $N_{12}$ | $\cdots$ | $N_{1s}$ | $R_1$ |
| Category 2 | $N_{21}$ | $N_{22}$ | $\cdots$ | $N_{2s}$ | $R_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| Category t | $N_{t1}$ | $N_{t2}$ | $\cdots$ | $N_{ts}$ | $R_t$ |
| Sum | $C_1$ | $C_2$ | $\cdots$ | $C_s$ | $N$ |

The $p - value$ of FET is computed by the following formula:

$$p = \frac{(\prod_{i=1}^{t}(R_i))(\prod_{i=1}^{s}(C_i))}{N! \cdot \prod_{i=1,j=1}^{t,s}(N_{ij}!)} \tag{10}$$

# B Memory Partition in Intel SGX

SGX partitions the main memory in two regions: protected region and unprotected region (as shown in Figure 1). Unprotected memory region is generally accessible. However, access to protected memory region is restricted. Data used by enclaves are swapped in and out of a segment of protected memory region, which is called *Enclave Page Cache (EPC)*. Enclave pages are stored in the EPC. Size of an enclave page is 4KB.

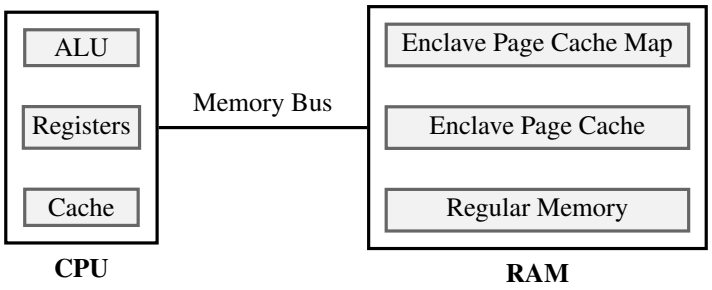


Figure 1: Memory partition in SGX.

Access to EPC is restricted on the hardware level. The *Enclave Page Cache Map (EPCM)* keeps track of which enclave is authorized to access which pages in the EPC. In EPCM, each page is mapped to a set of enclaves that are authorized to access it. SGX checks whether a specific enclave is authorized to access a page based on the EPCM.

# References


[1] Ali Shahbazi, Fattaneh Bayatbabolghani, and Marina Blanton. Private computation with genomic data for genome-wide association and linkage studies. In *International Workshop on Genomic Privacy and Security (GenoPri)*, 2016.

[2] Kristin Lauter, Adriana López-Alt, and Michael Naehrig. Private computation on encrypted genomic data. In *Progress in Cryptology-LATINCRYPT 2014*, pages 3–27. Springer, 2014.

[3] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.