Practical Privacy Attack on Aggregate Genomic Data

Moad Hani 1†, Ravindra Kumar2†, S. Varrette* and Frederic Pinel*
†*Univerity of Luxembourg, MICS*
**University of Luxembourg, FSTM*
2, avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg
Firstname.Name@uni.lu

***Abstract*—**

Genome-Wide Association Studies (GWAS) identify the genomic variations that are statistically associated with a particular disease in a population. Publicly sharing aggregate statistics collected by GWAS is foreseen to benefit the development of personalized medicine.

However, several attacks on the statistics computed by GWAS have been described in the literature, which led to those statistics not being publicly shared anymore. One possible attack on GWAS consists in recovering the genomes of the individuals who participated in the study based on the released statistics.

In a state-of-the-art publication, it has been argued that a GWAS can be considered safe if sufficiently enough genomes have been used (using an equation). However, this relies on a complexity argument, and not on a security proof.

The goal of this project is to experimentally challenge this statement. More particularly, we will start with a brute-force attack, which will then be refined (cutting the exploration tree) and parallelized (using multi-threading (MPI), GPU programming and the HPC facilities).

1. Introduction

Genome-wide association studies (GWAS) involve testing genetic variants across the genomes of many individuals to identify genotype–phenotype associations. GWAS have revolutionized the field of complex disease genetics over the past decade, providing numerous compelling associations for human complex traits and diseases. Despite clear successes in identifying novel disease susceptibility genes and biological pathways and in translating these findings into clinical care, GWAS have not been without controversy. Prominent criticisms include concerns that GWAS will eventually implicate the entire genome in disease predisposition and that most association signals reflect variants and genes with no direct biological relevance to disease. In this Review, we comprehensively assess the benefits and limitations of GWAS in human populations and discuss the relevance of performing more GWAS.

Single nucleotide polymorphisms (SNPs) are the most common type of variation in the human genome, and represent the substitution of a single base with another. At least 50 million SNPs have been identified in the human genome and they can be found across the entire genome. It is possible to look at these single points in the DNA code to see whether a person has a SNP in that position within their DNA. That's being said researchers found that we can still recover the (under certain conditions) the complete sequences of SNPs.

This paper is organized as follows:

After this brief introduction, section 2 will look at the key concepts and motivation behind this work, and then discuss the kind of GWAS attack we have been working on. This will be followed by a discussion of the technologies used to implement and a comparison of two approaches in section 3, the first being in brute force (sequential code) and the second more appropriate with a very short execution time. The discussion of results will be included in section 5 and the related work detailed in section 6. Finally, the conclusion is intended to give an idea of the progress of our work and to evoke some important points as well as to give food for thought on the related issues.

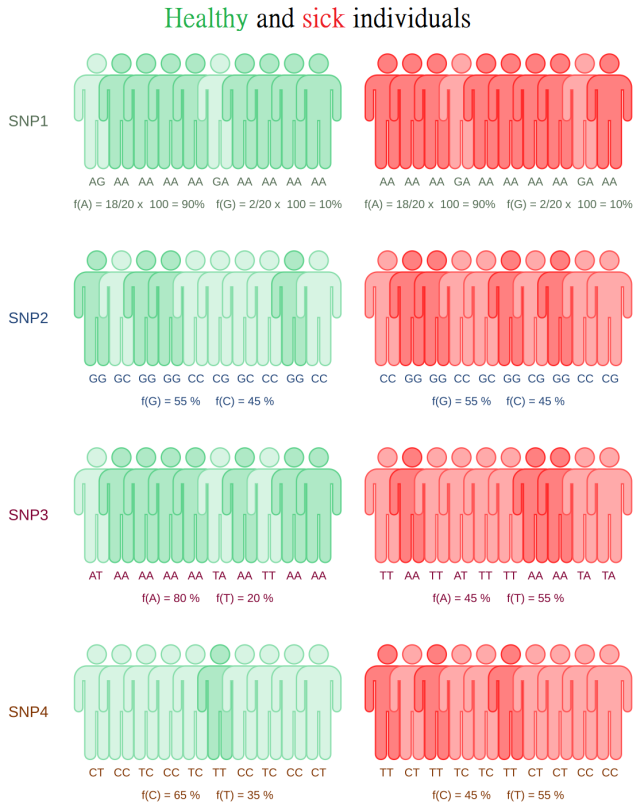
2. Context & Motivations

In contrast to rare monogenic diseases, there are frequent multifactorial diseases for which there is a certain degree of genetic predisposition. These may include diabetes, hypertension, myocardial infarction...

The hypothesis here is that certain alleles or combinations of alleles are not pathogenic per se, but predispose, promote, or facilitate the onset of disease. Their detection in an individual does not in any way mean that the disease will appear (contrary to monogenic diseases very often), but that the individual has a greater risk of presenting this type of pathology, which is otherwise frequent, and that it must be monitored, detected, and prevented.

Genome-wide association studies (GWAS) are used to highlight these contributing factors. The aim is to study the frequency of thousands of SNPs in large series of patients with the same frequent pathology, in comparison with a cohort of healthy patients who are comparable in all other respects (age, ethnicity, lifestyle, gender, etc.). In general, the result highlights the slightly more frequent association

of a common SNP with a common disease (promoting effect) or the opposite (protective effect). However, these studies do not make it possible to say whether it is the mutation of the SNP itself that promotes or protects against the disease, or whether it is a neighbouring mutation genetically linked to the SNP.



Whole genome association studies consist of sequencing the entire genome of several individuals, both healthy and sick. Here, for simplicity, only ten individuals of each type are represented. The millions of SNPs of these individuals are then compared. In the example presented here, only four SNPs are represented. SNP1 and SNP2 are as common in the healthy population as they are in the sick population. This means that the gene responsible for the disease is not located near these SNPs. Therefore, they are not of clinical interest. For these two SNPs, individuals carrying at least one copy of the less frequent allele are shown in light color, while homozygotes for the more frequent allele are shown in dark color.

For SNP3, the frequency of the T allele is higher in sick individuals than in healthy individuals. The gene causing the disease is therefore close to SNP3. Individuals carrying at least one copy of the T allele are shown in light color, the others in dark color.

Finally, the C allele of SNP 4 is more present in healthy individuals than in patients. This may mean that this version of the SNP is linked to an allele of a gene that has a protective effect against disease. Individuals carrying at least one copy of the C allele are shown as light-colored, the others as dark-colored.

For SNP3, it can be seen that most (light red individuals) but not all (dark red individuals) of the diseased individuals carry the T allele. Similarly, for SNP4, one of the healthy individuals does not have allele C (dark green individuals). This means that the disease studied here is not monogenic with complete penetrance. On the contrary, it is a multifactorial disease, with at least one gene, with incomplete penetrance, having a deleterious effect, and one gene having a protective effect.

Finally, it should be noted that, in the examples presented here, the percentages of the two alleles are either exactly the same or very different in healthy and sick individuals. In practice, the results obtained are not necessarily so extreme. It is therefore necessary to use statistical tests to determine the probability that the results obtained are due to chance or not.

The objectives of a genome-wide association study are:

- To scan several thousand SNPs on many individuals to find genetic variations associated with a particular disease.
- Help develop better strategies to detect, treat and prevent the disease.
- GWAS are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

GWAS are frequently used to map genotypes (the genes within an organism) onto phenotypes (the traits of an organism) [1]. The predominant application for GWAS today is in the study of genetic diseases. GWAS are conducted by examining genetic mutations which differ significantly between individuals who have an illness and those who do not. These individuals are partitioned into the case and control groups, respectively. What follows are brief descriptions of some of the relevant terms in genetics, as well as the statistics we would like to compute over the input data.

SNPs In genetics, a DNA sequence consists of multiple nucleotides, where a single nucleotide can take one of four values 'A', 'G', 'T' and 'C'. A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation in which a single nucleotide varies between individuals in a population. Given that nearly 99% human DNA are identical, the study of identifying genetic mutations such as SNPs is essential in determining which genotypes correspond to which human traits. During the past decade, the associations between a number of common diseases (e.g., heart disease, diabetes etc.) and common SNPs have been widely studied [1].

Minor allele frequency (MAF): An allele is a variant of the same gene or the same genetic locus. A minor allele frequency (MAF) is the frequency at which the least frequent allele occurs [1] within a given population. For example, the genotypes of five individuals at the same loci are as follows, AA, AG, AA, AG, and GG. Since G is less frequent than A, and its frequency is 410, The MAF can be calculated as 0.4 in this case.

χ^2 statistic: The χ^2 statistic is the statistic used by a χ^2 hypothesis test. Given a set of categories and the frequencies with which observed and expected values fall into those categories, a χ^2 test can be used to test whether the observed and expected populations differ in a statistically significant way. The χ^2 statistic is computed as $\sum_{i,j=1}^{n} \frac{(Obs_{i,j} - Exp_{i,j})^2}{Exp_{i,j}}$ (1)

where $Obs_{i,j}$ and $Exp_{i,j}$ denote the observed and expected allele counts from allele type j (e.g., j ∈ A,G in above example) in group i(i ∈ case and control groups).

Difference between linkage and association Linkage studies

- Collect set of families with individuals carrying disease or phenotype
- Look for co-segregation of small number of markers with disease status.
- Relies on inheritance of large segments of DNA seperated by a few genetic recombinations

Association Studies

- Collect unrelated individuals and look at allele frequency differences between cases and controls (or cases and parents for TDT)
- Requires genotyping hundreds of thousands of markers.
- Relies on the many unobserved recombinations leaving correlations between nearby genetic variants along chromosomes within the population

Attack on GWAS

Because the results of the GWAS study are statistical in nature, most researchers thought until recently that it was safe to share and publish such depersonalised results. This belief has been challenged by recent bioinformatics research. In this work we're interested in the recovering attack that claims given certain conditions we can fully or partially recover all the elements of the matrix by knowing the public release.

3. Implementation and Experimental Setup

Here are the technologies we used and the steps we follow for our implementation:

4. The difference between the two approaches

The key objective is to recover all elements of a given matrix and generate the associate permutations.

4.1. Sequential Approach

- Read Input
- Compute the sum of each column

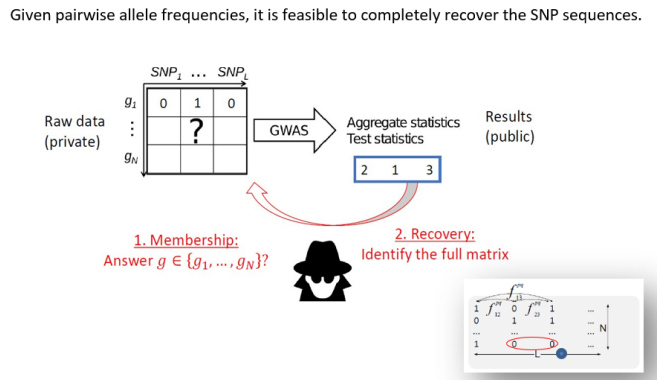


Figure 1. Illustration of the method followed by an interceptor to recover the SNP sequences

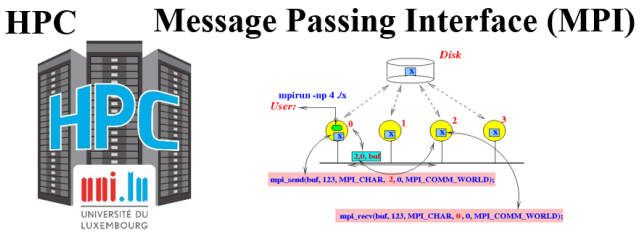


Figure 2. We use HPC facilities and the MPI for our experimentations

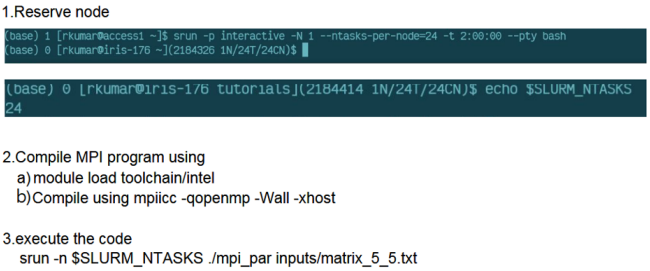


Figure 3. The three main steps we followed to explore the solutions

- Send each column sum and row number to twiddle function and then permutes the elements and store into columns, juxtapose these columns and print matrices.

Coded by Phillip J. chase (Algorithm 382 combination of M out of N objects).

4.2. Our Approach

We have one main process and all the rest are workers processes.

4.2.1. Main Process.

- Read all inputs
- Calculate all the sum column wise

- prepare tidlle data (for first matrix, we just sort according to column size)
- take all permutation in column one
- Spread these permutation through the workers

4.2.2. Worker Process.

- Based on the received data start calculating all the possible permutation with given option for the column 1 and return results.
- If any matrix fills all the requirements it is printed.
- When all permutations are tested it returns the possible solutions count with the given first column.

5. Validation and Experimental Results

This section presents the results obtained. All computation operations were done using HPC and parallelism using MPI. This table shows the results

Matrix Size	Total number of solutions	Sequential code time (in seconds)	Parller code time (in seconds)
3,3	2	0.000	0.000
4,4	15	0.000	0.000
5,5	483	0.010	0.000
6,6	41280	1.680	0.530
7,7	2798813	607.460	63.030

Figure 4. Table - Summary of results

obtained by comparing the sequential time and the time of the parallel calculation.

While the former grows exponentially the latter always remains within the tolerable norms and this is the importance of distributed computing over several workers (secondary nodes) and shows all the interest of our work.

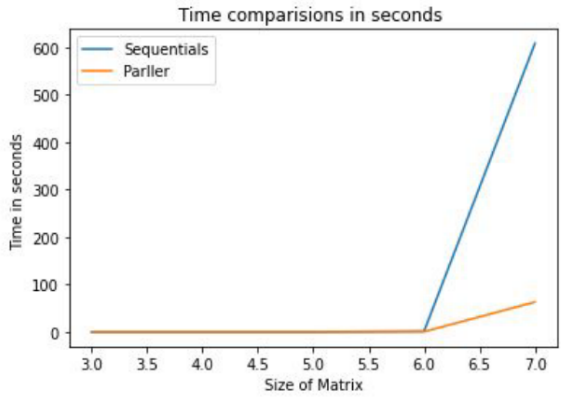


Figure 5. Time comparison related to the Sequential and Parallelized code

The generated matrix has a size of 8*8 and is the last of its kind as it is due to the server release time. So unfortunately our session is finished before we get all the results for the 8*8 matrices. Solution 73. 413. 147 was therefore the last one we were able to generate. This shows the constraint in the genomics of computing during all the GWAS workflow as we explained during our presentation of this work.

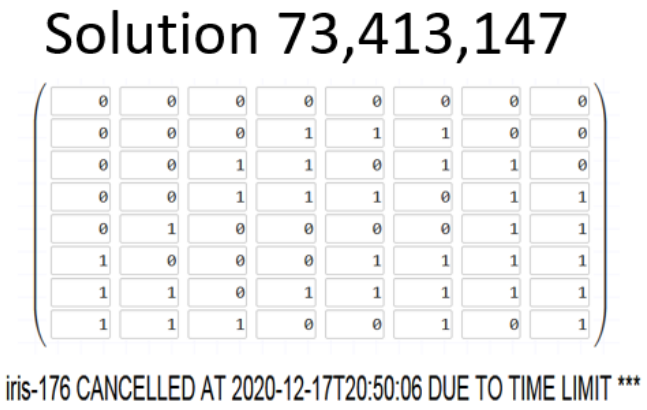


Figure 6. Final Matrix - Due to time out

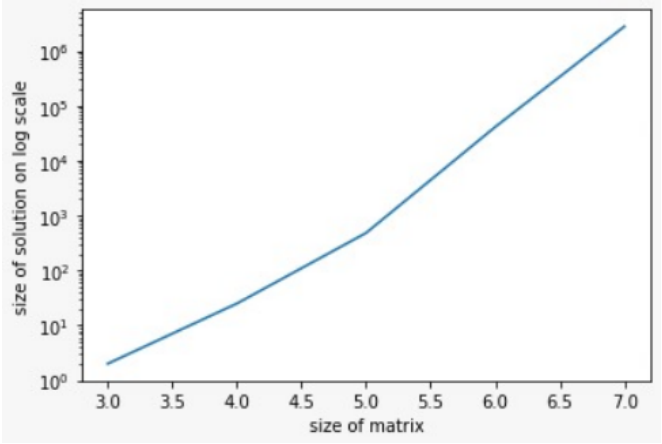


Figure 7. Size of solutions in log scale

6. Related Work

Our work and experimentation are based on The HPC facilities [2].

Indeed many papers discussed the concepts of GWAS and its utility [3] [4] [5] [6] [7] and the recent progress in human genome research has made a great demand on convenient access to sensitive human genome data for research purpose. The problem of balancing privacy protection and data sharing in this domain, however, has not been seriously studied until Homer, et al. published their findings a couple years ago. After that, several research groups, including us, have started working on this important issue. As a prominent example, Sankararaman, et al recently propose a technique

(SecureGenome) for measuring the maximum statistical powers achievable on a set of single-allele frequencies. Most of these studies focus on single allele frequencies, which has been found in prior research to be insufficient, as sensitive information can also be inferred from other sources like test statistics. Our work research is an attempt to understand and assess the computation side under typical inference threats.

Homomorphic encryption approach

There has already been some work on using homomorphic encryption to preserve the privacy of the patients while performing statistics on genome data. Kim et al. [8] present the computation of minor allele frequencies, and the χ^2 statistic with the use of the homomorphic BGV and YASHE encryption schemes. They use a specific encoding technique to improve on the work of Lauter et al. [5]. However, they only compute the allele counts homomorphically, and execute the other operations on the decrypted data. Another work on GWASs using fully homomorphic encryption was published by Lu et al. [9]. They also start from encrypted genotype/phenotype information that is uploaded to a cloud for each person separately. Then they perform the minimal operations necessary to provide someone with access to the decryption key with the necessary values to construct the contingency table for the requested case based on the data present on the cloud. Hence, when performing a request, the scientist gets three encrypted values, and based on those he can, after decryption, reconstruct the contingency table, and compute the χ^2 statistic in the clear. These solutions are not resistant to attacks like the one described by Homer et al. [10].

Sadat et al. [11] propose a hybrid system called SAFETY, to compute various statistical values over genomic data. This hybrid system consists of a combination of the partially homomorphic Paillier scheme with the secure hardware component of Intel Software Guard Extensions (Intel SGX) to ensure both high efficiency, and privacy. With this hybrid system they propose a more efficient way to get the total counts of all patients for a specific case. By using the additive property of the homomorphic Paillier scheme, they reduce the computational overhead of decrypting all individual encrypted outputs received from the different servers. Afterwards it uses the Intel SGX component to perform the χ^2 computations. Even though, the results of this system scale well for increasing number of servers that provide data for the computation. Sadat et al. [5] mention that the only privacy guarantee for the final computation result against the attack described by Homer et al. [10] is the assumption that the researcher decrypting the result is semi-honest.

Zhang et al. [12], construct an algorithm, which performs the whole χ^2 statistic in the homomorphic domain. To compute the division, they construct a lookup table in which they link the result of their computation with the nominator and denominator of the corresponding, simplified fraction. Therefore, an authenticated user can look up the correct fraction in the lookup table after decrypting the result, and hence recover the result of the χ^2 statistic. Even though their strategy performs well, it does not scale enough to treat the large datasets we envision in our application. Increasing the number of patients in the study would increase the circuit depth significantly, which comes with several disadvantages including increasing the parameter sizes, and hence the key size, and ciphertexts size, as well as the computation time.

Secure multiparty computation approach

Kamm et al. [13] propose a solution to address the privacy challenges in genome-wide association studies. Their application scenarios, much like ours, focus on large data collections from several biobanks, and their solutions are based on the same fundamental techniques as ours. However, the setting of Kamm et al. [13] requires all raw genotype, phenotype, and clinical data to be entered to the secure shared database. To the contrary, our setting assumes that only the aggregate values, necessary to identify the significance of a gene-disease relationship (i.e., the contingency tables recording the counts of genotypes vs. phenotypes), are contributed by each biobank. This is a simpler, and more realistic setting, which not only is likely to be implemented in the near future, but also alleviates the computational cost of the proposed solutions. Unlike the approach of Kamm et al. [13], and the alternatives that they suggest.

Independent and concurrent work by Cho et al. [14] tries to address the same problem as we do in our work, using multiparty computation techniques. They focus on a method that enables the identification and correction for population biases before computing the statistics. However, just like the work of Kamm et al. [13], they make the strong assumption of semi-honest security. In practice, the semi-honest security is not a sufficient security guarantee for GWAS, as attackers who have obtained access to the systems are likely to employ active measures to obtain the data.

Constable et al. [15] present a garbled-circuit based MPC approach to perform GWAS. Their solution can compute in a privacy-preserving manner the minor allele frequency (MAF), and the χ^2 statistic. Similarly to the work of Kamm et al. [13], the framework of Constable et al. [15] requires the raw genotype, and phenotype data, increasing the workload of the proposed privacy-preserving system. The solution of Constable et al. [15] only works for two medical centers. Despite the strong security guarantees that our approach offers, which generally presents itself as a tradeoff to efficiency, our proposal is faster than that of Constable et al. [15]. This is also due to the fact that we have optimized the computations of the χ^2 statistic, in such a way that the expensive computations in the privacy-preserving domain, are avoided to the maximum extent possible.

Zhang et al. [16] propose a secret-sharing based MPC approach to solve the same GWAS problem as Constable et al. [15]. Although Zhang et al.'s solution can scale to more than two medical centers contributing data to the GWAS, the approach has the same inherent limitations (e.g., requiring raw genomic data as input) that their application scenario incurs. The works of Zhang et al. [16], Constable et al.

[15], and Cho et al. [14] have not considered protecting the aggregate statistic result of the private computation, which –as Homer et al. [10] showed– can be used to breach an individual's privacy. We additionally protect the aggregate statistic result, while at the same time allowing for a public list to be created, showing which SNVs are significant for a certain disease.

7. Conclusion

Genome-wide association studies (GWAS) aim at discovering the association between genetic variations, particularly single-nucleotide polymorphism (SNP), and common diseases, which have been well recognized to be one of the most important and active areas in biomedical research. Also renowned is the privacy implication of such studies, which has been brought into the limelight by the recent attack proposed by Homer et al. Homer's attack demonstrates that it is possible to identify a participant of a GWAS from analyzing the allele frequencies of a large number of SNPs.

Recovering SNP sequences is related to the research on contingency table release, and discrete tomography, which tries to reconstruct a matrix from a small number of projections. However, the specific problem of restoring a matrix from pair-wise allele counts is new and the related complexity of the problems is very important to be studied in the afferent paper (To release or not to release).

Through this work we have been able to observe the importance of parallelism in genomics but also the related constraints. Since some limits cannot be exceeded even with a very powerful supercomputer such as HPC. In addition, we became familiar with MPI and OpenMP in different environments, such as : Linux, Windows (locally) and of course the IRIS Cluster where we launched jobs of 3 successive hours to get the results.

Acknowledgments

We would like to heartily thank our supervisor Dr. Jérémie DECOUCHANT for his valuable time, availability and generous guidance.

The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg [2]

References

[1] M. M. Y. J. A. J. H. G. Visscher PM, Brown MA, "Five years of GWAS Discovery," Jan 2012, pp. 90(1):7–24.

[2] S. Varrette, P. Bouvry, H. Cartiaux, and F. Georgatos, "Management of an Academic HPC Cluster: The UL Experience," in *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*. Bologna, Italy: IEEE, July 2014, pp. 959–967.

[3] S. et al., "Genome-wide association study identifies five susceptibility loci for glioma. Nature Genetics," 2009, pp. 41 (8) : 899–407.

[4] S. e. a. Sanson*, Hosking*, "Chromosome (1) - Variation influences glioma risk. Human Molecular Genetics (2)," 2009, pp. 1–7p11.2 and 2–15;20(14):2897–904.

[5] Wang and Shete, "Using Both Cases and Controls for Testing Hardy-Weinberg Proportions in a Genetic Association Study. Human Heredity," 2010, p. 69:212–218.

[6] ——, "A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study BMC Genetics," 2011, p. 12:3.

[7] S. et al., "Genome-wide high-density SNP linkage search for glioma susceptibility loci: results from the Gliogene Consortium. Cancer Research," 2012, pp. 71(24) 7568–7575.

[8] L. K. Kim M, "Private Genome Analysis through Homomorphic Encryption." 2015, p. 15:3.

[9] S. J. Lu W. J, Yamada Y, "Privacy-Preserving Genome-Wide Association Studies on Cloud Environment using Fully Homomorphic Encryption." 2015, p. 15:1.

[10] R. M. D. D. T. W. M. J. P. J. S. D. N. S. C. D. Homer N, Szelinger S, "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures using High-Density SNP Genotyping Microarrays. PLoS Genet." 2018, p. 4(8):1000167.

[11] M. N. C. F. W. S. J. X. Sadat MN, Aziz MMA, "SAFETY: Secure GWAS in Federated Environment through a Hybrid Solution with Intel SGX and Homomorphic Encryption." 2017, p. 1703.02577. [Online]. Available: http://dblp.uni-trier.de/rec/bib/journals/corr/SadatAMCWJ17

[12] J. X. X. H. W. S. Zhang Y, Dai W, " FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption. BMC Med Inform Dec Making." 2015, p. 15(5):S5. [Online]. Available: https://doi.org/10.1186/1472-6947-15-S5-S5

[13] L. S. V. J. Kamm L, Bogdanov D, "A New Way to Protect Privacy in Large-Scale Genome-Wide Association Studies. Bioinformatics." 2013, p. 29(7):886–93.

[14] B. B. Cho H, Wu DJ, "Secure genome-wide association analysis using multiparty computation. Nat Biotechnol." 2018, p. 36(6):547.

[15] W. S. J. X. C. S. Constable SD, Tang Y, "Privacy-Preserving GWAS Analysis on Federated Genomic Datasets. BMC Med Inform Decis Mak." 2015, p. 15(5):2.

[16] A. G. Zhang Y, Blanton M, "Secure Distributed Genome Analysis for GWAS and Sequence Comparison Computation. BMC Med Inform Decis Mak." 2015, p. 15(5):4.