# Investigating Privacy Preserving Technique for Genome Data

Slesha S. Sanghvi(✉) and Sankita J. Patel

Department of Computer Engineering, Sardar Vallabhbhai National Institute of Technology, Surat 395007, Gujarat, India
slesha07@gmail.com, sankitapatel@gmail.com

**Abstract.** The rapidly growing genome sequencing technology has enabled the production of huge amount of sensitive genomic data. Presently a-days, it is conceivable to create highly detailed genotypes at lower cost. Sharing of genomic dataset is a key to comprehend the hereditary premise of human ailments. Because of the sharing of such information, genuine privacy challenges emerge with the expanded number of hereditary tests and immense gathering of such genomic information. The expanded accessibility of such information has real ramifications for individual protection, since it contains basic elements of human as well as contains, illnesses points of interest, insights about relatives, past and future era, responses to medication and substantially more.

To overcome the privacy issue in genomic data, previously some solutions had been purposed based on encryption techniques. However, the existing solutions has some limitations viz., identification of an individual from Genome Wide Association Study (GWAS) sets, generated test results contain Single Nucleotide Polymorphism (SNP) information about patients etc. In this work, we aim to propose a privacy preserving technique for genomic data that strengthen the security of genomic data.

**Keywords:** Genome sequencing · Privacy preserving · Disease detection
Privacy preservation for genome · Medical data security

## 1 Introduction

Genome fundamentally represents blueprint of a body. Our appearance, our maladies related data, our family history and much more information is determined by genome. Major applications of genome processing include recognizing criminals, prenatal testing and premature identification of diseases. Physically, it would require lots of time and efforts to establish the correlation between genomic information and human characteristics i.e. eye shading, obesity and so forth [1]. The healthcare can be revolutionized by medical data based on genome; however, the genomic data is also vulnerable against mishandle at the exact time. Essentially, genome data leads to social stigma, discrimination, employment and insurance denial. Human genome can put lifelong impact on an individual's life if it is leaked, as it is particularly stable. In literature, there exist numerous attempts that identify the risk of publishing genomic data [2, 3]. In Homer et al. [2], authors demonstrated that an individual could be


Z. Patel and S. Gupta (Eds.): ICFITT 2017, LNICST 220, pp. 106–115, 2018.
https://doi.org/10.1007/978-3-319-73712-6_11

identified by using aggregate genomic data. Therefore, the data usage management of genomic data is crucial.

Genomes are basically acquired from a person in chemical form for digitization prepare. Both forms of genome i.e. digital and chemical; should be anticipated as it can be misused by some adversary. Due to inappropriate management, privacy breaching of genomic data can happen which leads to identity risk of individual. Therefore, genomic data must be handled carefully [4].

### 1.1 Motivation

Larger part of health care services requires genomic data to perform productive medical research or to perform certain diseases susceptibility test. This in turn requires sharing of genomic data to researcher or some third-party agency.

In addition, genetic data sharing among hospitals and research institutions is imperative for large-scale genetic studies. For example, let us consider that two medical organizations own the genetic datasets of their patients. The organizations need to run machine learning algorithms on the union of the datasets they own, without revealing their datasets to each other. Without utilizing a safe convention for these organizations to share data joint calculation on this information is infeasible. Other problem includes secure protocols for individual patient's disease susceptibility tests.

To address this issue, some previous work has been done, especially by Jha et al. [5]. Authors in [5] explore privacy preserving analysis for personal genomics. The idea is to utilize the outcomes found in Genome Wide Association Studies (GWAS) basically, to examine a particular disease susceptibility of an individual for getting a specific disease based on certain genetic markers that includes allele frequency and molecular markers. Limitation of this approach is mainly, it does not contain any secure way for examining an individual.

### 1.2 Contribution

As discussed, we focus on the problem of privacy preservation in Genome dataset. The problem is described as below:

> *Suppose, there is a data provider owing a private Genome dataset D. The dataset is required to be shared with various organizations for the two purposes viz. (1) medical research and (2) disease susceptibility test. The goal is to preserve the privacy of an individual, whose genome sequence is stored in D, while maintaining the accuracy of the results.*

To address the problem, we proposed a technique to preserve the privacy of an individual using Paillier cryptosystem and differential privacy. By using this technique an individual can generate his/her test results in secure way. Generated test results are then used by researchers to get statistical information without breaching individual's privacy.

In upcoming Sect. 2 we discuss about genomic background and issues related to securing genomic data. After that in next sections our proposed technique and its performance results are described.

## 2 Background

The genome of a human body contains set of genetic data that consist of mainly four different bases viz. Thymine (T), Guanine (G), Adenine (A), and Cytosine (C). A chromosome contains genetic data and these genes are accountable from form of functions dominant in human body all at once. There is distinction within the arrangement of those bases, which are supported by every DNA strand that results in individuation between individuals' genetic composition. Due to genetic differences, DNA of every person is different from reference genome by approximately 0.5%. SNP (Single nucleotide Polymorphisms) of a human body is commonest genetic dissimilarities [6].

### 2.1 DNA Sequencing and Analysis Process

DNA of the individual is collected from varied sample sources viz. skin, hair, saliva and blood. Once sample is collected, using extraction kit of DNA genetic data is extracted and then the process of sequencing that data is started using any sequencing platform. One of the widely used sequencing platforms is Illumina Sequencer [6]. For standard bioinformatics analysis, the digital DNA data is used after the sequencing process of DNA. In this manner, just physical securities are not sufficient to shield protection and supply wellbeing as computerized information is regularly replicated, changed and shared [6].

### 2.2 Features of Genome

After discussing about how genome information is handled, given us a chance to examine why genome data is delicate, and require more privacy than the standard medical data. Following are the major features of genomic data that creates privacy issues:

Genome contains sensitive information, which may bring about separation, work refusal and protection dissent, and mortification [4]. Immaterial intergenerational change i.e. DNA of individual changes less from one era to future era [7]. Likeness with blood-relatives i.e. one human genome contains loads of sensitive data regarding his blood-relatives. Closely connected peoples have very alike genomes [8]. Information contained by genome has various applications containing biomedical analysis, healthcare etc. [4]. By using partial genomic data, it is possible to get unavailable data i.e. it can leak disease information which is not even available. Human hereditary data contains six billion nucleotides which is very large in size.

### 2.3 Privacy Breach in Genome Data

Privacy breaching techniques of genomics are of three types:

**Identity Tracing**
In this kind of attack, intruder can build up a linkage between the data owner's hidden identity and an unidentified genome.

**Attribute Disclosure Attack via DNA (ADAD)**
Access of distinguished DNA by intruder and furthermore without utilizing express identifiers database that interfaces delicate properties with DNA-inferred information. These methods look at DNA information and give interface between the target's personality and its sensitive quality.

**Completion technique**
Intruder has just access of sanitized dataset without knowing about delicate area. In addition, intruder knows personality of genomic dataset however no get to. Point is to reveal the delicate area i.e. not a piece of genuine information [9].

## 2.4 Privacy Issues in Genome Data

Digital genomic data are used in various bioinformatics processes viz. searching on a genetic dataset, querying private data of genome, and sequence alignment. Regardless of their helpfulness in the medical field, these procedures present high danger of leakage of private data.

In first issue, insecure environment process of sequence alignment, DNA sequence alignment process demands high and expensive computation which is outsource to publicly available clouds. Yet, sending such kind of personal information to an open cloud may raise an issue as it is controlled by some third-party associations which create privacy issue [6].

In second issue, querying private data related to genome, as discussed, the human genome contains private data around an individual's science like whether an individual has a probability to build up a particular kind of disease. For the prevention of disease and furthermore for prescribing customized medicine, person's genome is taken and used to query against a list of known variations of disease to calculate susceptibility of diseases. In addition, to decide biological connections between persons, it is required to query genomic data to get result of tests like Paternity and ancestry [6].

In third issue, Private Data Sharing for GWAS, Genome-Wide Association Study (GWAS) defines connection between specific traits and common variations of genetics. From genomic records of thousands of individuals, GWAS examines Single Nucleotide Polymorphisms (SNP) and then it produces aggregate statistics. These aggregate statistics is then used to find connection between a disease and a SNP [10]. Homer et al. [2] presented an outline for robust and accurate detection of the existence of an individual by some known genotype in the mixture of complex DNA. The individual distance is measured from a test population and a reference. Then based on it t-test is calculated by using previously unknown individuals and the distance metrics to analyze this two populations and get the difference between them.

## 2.5 Related Work

As mentioned earlier for several types of privacy issues there are different techniques applied on genome data, but as shown in Table 1 all techniques having some of limitations. So, to improvise privacy for genomic data we proposed technique to resolve the above-mentioned issues.

**Table 1.** State of art in literature in privacy preserving techniques for Genome data

| Privacy issues | Citation | Proposed solution | Limitations |
|---|---|---|---|
| Secure alignment on insecure environment | Chen et al. [11] | Use of seed-and-extend method | Higher data volume and high computational power |
| Querying private data related to genome | Alday et al. [12] | Use of Homomorphic encryption | Generated genetic tests results contains overall information regarding SNPs |
| Privacy preserving data sharing for GWAS | Fienberg et al. [10] | Uses ε-differentially mechanism | Releasing summary statistics for large data sets may not be enough to ensure the privacy of individuals |

## 3 Proposed Technique

As discussed, we aim to preserve privacy for genome data. In the proposed technique, we utilize two significant methods, viz., Paillier cryptosystem [13] and differential privacy [14, 15]. The block diagram of our proposed approach is shown in Fig. 1.

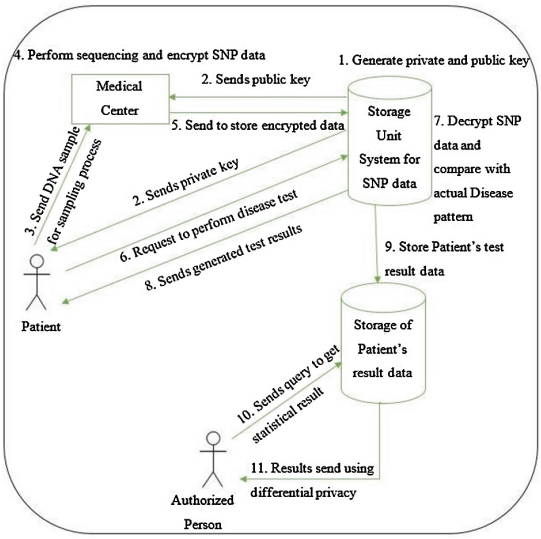


**Fig. 1.** Proposed technique

As shown in the Fig. 1, whenever patient enrolls for DNA sequencing process, patient gets private key. Using private key, patient will generate his/her test report by directly contacting the system. The generated results are then stored for statistical data analysis.

### 3.1 Generation of Secure SNP Data and Statistical Information

To achieve security, we apply Paillier cryptosystem on SNP data of the patient. Whenever patient enroll for DNA sequencing process at that time Storage Unit System (SUS) generates private and public key for the patient using Paillier cryptosystem. After the sequencing process, which is done at either medical center itself or by some third party, they will encrypt the SNP data of patient using patient's public key and store encrypted SNP data file of patient in SUS. If patient wants to go for any particular disease test, he/she will simply request to system. Patient forwards his/her private key to SUS. At SUS, system will perform the decryption operation. After decrypting the file, that file will be compared with particular disease's SNP file. For these tests, there are so many number of pattern files is stored at SUS. In this file, there is SNP information, which shows particular disease's pattern. By comparing this disease pattern file with patient's file test results are generated. After generation of test result, generated decrypted file of patient's information will be deleted automatically by the SUS. Moreover, generated test result will be forwarded to the patient.

Generated test results are stored in a file, which contains the basic information of patient i.e. patient id, sex, age and test result. This file will be further used for research purpose where researchers can send the queries to get some statistical information from database like, "How many number of patients having breast cancer who are male?" However, as explained before this generated result of query is prone to disclose the identity of patient. So, for statistical results we used differential privacy. Using Laplacian noise, we add noise into generated query result. After generating different statistical results, we add homomorphic encryption on these results by using Paillier cryptosystem.

## 4 Experimental Setup

In this section, we are going to discuss about experimental setup that we have created for Paillier cryptosystem and differential privacy. We have implemented Paillier cryptosystem in JAVA programming language and for differential privacy we have used R tool.

### 4.1 Paillier Cryptosystem Setup

Parameters to be used for key generation: We have taken public and private key of the length of 512 bits. To generate two random prime numbers p and q we have used 256 bits of length with certainty of 64 and used Random () function, this shows that randomly generated prime numbers are positive Big Integers that is probably prime with the length of 256 bits. Using p and q, we have generated n which is of 512 bits. For public key one more parameter is needed i.e. g. g is generated randomly using random function in the class of $Z^{*2}_{n}$ of 512 bits. Based on p and q, lambda is generated. u is generated using g, n and lambda. We generated u direct at the time of decryption. So, patient having private key as lambda only.

### 4.2 Differential Privacy Setup

To preserve privacy in statistical results, we should add noise. So, for that ε value should be set as small as possible to get privacy of statistical results. On selected dataset, we perform number of cycle to get the value of ε at which we can get very minimum difference in between original dataset and by changing one of the row values of the dataset. We set threshold value as sum of all detected disease column. Number of time cycle we run, we set it as 1000. We took ε value in between 0 to 1 with the increment of 0.01. For adding noise, we have taken one parameter named as sigma who indicates $\Delta f/\varepsilon$ and its value is $1/\varepsilon$. As for our dataset sensitivity function carries value 1, which means $\Delta f$ contains value as 1. We add Laplace noise over here as Lap $(1/\varepsilon)$. Also, we created one bound over here as added noise will be in a boundary of 0 to 2.

### 4.3 Dataset to Be Used

For genomic dataset, we take dataset from GWASdb. GWASdb is a database of SNP-phenotype association mapped to Disease Ontology and Human Phenotype Ontology [16]. These datasets basically contain the disease information related to chromosomes which affects DNA. These datasets stored at SUS, which will be used at the time of susceptibility testing process to compare it with patient's chromosomes.

After completing susceptibility test process generated dataset contains basic information related to patient i.e. patientID, sex, age, disease detected etc. For this we had taken dataset from UCI machine learning repository. Over here we used heart disease dataset of Hungarian data taken from this repository [17].

## 5 Performance Results

Paillier cryptography was performed on SNP data of patient to generate susceptibility test. The implementation of this cryptosystem was carried out on Eclipse IDE. For this we have taken heart disease chromosome details file from GWASdb which contains 11805 chromosomes. For encryption of patient's data, it takes approximately 201912 ms. And for susceptibility test checking it takes approximately 399764 ms in average case where no need to compare whole file for test results. In between we get the result as test is negative. But for worst case scenario where test results come as positive it takes approximately 990835 ms. Decryption operation is also performed at the time of susceptibility test process.

After susceptibility test, generated results are used for the purpose of research. On this dataset we fired below queries,

1. How many numbers of patients are having disease?
2. How many numbers of patients are having disease who are male?
3. How many numbers of patients are having disease who are female?
4. How many numbers of patients are having disease whose age group is between min and max (as per user enters)?
5. How many numbers of patients are having disease whose age group is between min and max (as per user enters) and gender (male or female)?

Generated ε values for above queries are shown in below Fig. 2, 3 and 4.

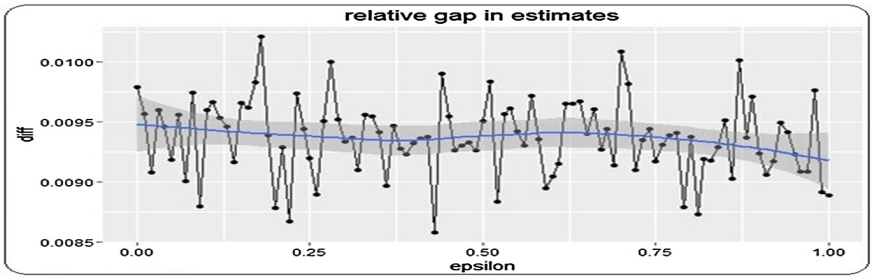


**Fig. 2.** Same graph for query 1, query 3, query 4 having age group 50–55 and query 5 with age group 50–55, gender female with ε value 0.43

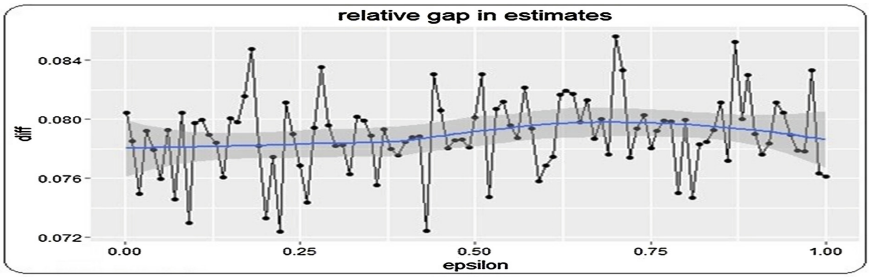


**Fig. 3.** Graph for query 2 with ε value 0.22

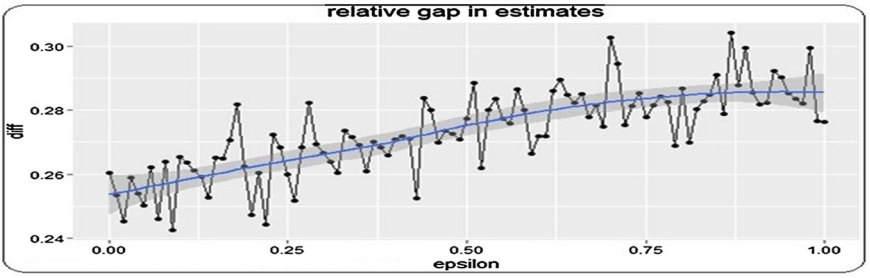


**Fig. 4.** Graph for query 5 having age group 50–55 and gender male with ε value 0.09

# 6 Result Analysis

## 6.1 Susceptibility Test Result Analysis

For susceptibility test of patients, we are having large amount of dataset so it takes more time for generating results. For comparison in between disease's chromosome file and patient's chromosome file its time complexity in best case is O(1) and in worst case scenario $O(n^2)$.

## 6.2 Statistical Result Analysis

As seen in the Figs. 2, 3 and 4, for epsilon values 0.43 we are getting lower difference. So, for this dataset of heart disease we can take ε as 0.43. In Fig. 3 we are getting two nearer values for ε i.e. 0.22 and 0.43, so we can take 0.43. But in Fig. 4 where our input is age in between 50 to 55 and gender having male we are getting privacy using very lower ε value that is 0.09. The actual result of this query is very low as, total number of patients having heart disease in between group of 50 to 55 and gender having male are only 3. So, from such kind of results, we can say that for different queries we are getting almost same values but for queries where we get very low population we got very less value of epsilon.

# 7 Conclusion and Future Work

As every year, bioinformatics field and sequencing process are becoming very important with increasing number of genomic tests. Persons use sequencing of DNA data for different number of aims and unwanted access to this sensible genetic information may create serious privacy breaching in the coming years. Due to the lack of techniques for privacy preservation it creates difficulties rather than benefits as there is revolutionary use of sequencing technology by medicine and health sciences.

With an aim to solve the privacy issue in genome data, in this paper, we mentioned the sensitivity of genome data and discussed various privacy breach techniques on genomic data. We proposed a technique based on the differential privacy and Paillier cryptosystem and discussed respective results.

In our work, after completion of susceptibility test there is no provision of encryption on generated result file. Future work could be to add encryption in result. The current work checks for a particular disease only that is, "Whether patient having X disease or not?" In future work, more feature can be added, using which, we can check and generate files containing details related to total number of chromosomes affected by particular disease. And then that statistical data would be store in a file with the use of differential privacy. And after performing susceptibility test patient can get the result as which diseases can be affected to them.

Major limitation of the proposed technique is the requirement of high computing power because there is a need for high number of chromosomes to be encrypted, decrypted and compared.

## References


1. Genome-wide association studies. http://www.genome.gov/20019523. Accessed 10 June 2016
2. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. **4**(8), 1000167 (2008)
3. Wang, R., Li, Y.F., Wang, X., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: CCS, pp. 534–544 (2009)
4. Naveed, M.: Hurdles for genomic data usage management. In: IEEE Workshop on Data Usage Management (DUMA), pp. 44–48, May 2014
5. Jha, S., Kruger, L., Shmatikov, V.: Towards practical privacy for genomic computation. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy, pp. 216–230 (2008)
6. Akgün, M., Bayrak, A.O., Ozer, B., Sağıroğlu, M.Ş.: Privacy preserving processing of genomic data: a survey. J. Biomed. Inform. **56**, 103–111 (2015)
7. Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al.: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science **328**(5978), 636–639 (2010)
8. Burdick, J.T., Chen, W.-M., Abecasis, G.R., Cheung, V.G.: In silico method for inferring genotypes in pedigrees. Nat. Genet. **38**(9), 1002–1004 (2006)
9. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**, 409–421 (2014)
10. Yu, F., Fienberg, S.E., Slavkovic, A.B., Uhler, C.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. J. Biomed. Inform. **50**, 133–141 (2014)
11. Chen, Y., Peng, B., Wang, X., Tang, H.: Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In: NDSS (2012)
12. Ayday, E., Raisaro, J.L., Hubaux, J.-P.: Privacy-Enhancing Technologies for Medical Tests Using Genomic Data, Technical report (2012)
13. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48910-X_16
14. Dwork, C.: Differential privacy. In: 33rd International Colloquium, ICALP 2006, Venice, Italy, Proceedings, Part II, 10–14 July 2006
15. Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., Naor, M.: Our data, ourselves: privacy via distributed noise generation. In: Vaudenay, S. (ed.) EUROCRYPT 2006. LNCS, vol. 4004, pp. 486–503. Springer, Heidelberg (2006). https://doi.org/10.1007/11761679_29
16. GWASdb SNP-Disease Associations dataset. http://amp.pharm.mssm.edu/Harmonizome/dataset/GWASdb+SNP-Disease+Associations. Accessed 10 June 2016
17. UCI machine learning database. http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.hungarian.data. Accessed 10 June 2016