# (Position Paper) Can You Really Anonymize the Donors of Genomic Data in Today's Digital World?

Mohammed Alser, Nour Almadhoun, Azita Nouri, Can Alkan, and Erman Ayday

Computer Engineering Department, Bilkent University, 06800 Bilkent, Ankara, Turkey

**Abstract.** The rapid progress in genome sequencing technologies leads to availability of high amounts of genomic data. Accelerating the pace of biomedical breakthroughs and discoveries necessitates not only collecting millions of genetic samples but also granting open access to genetic databases. However, one growing concern is the ability to protect the privacy of sensitive information and its owner. In this work, we survey a wide spectrum of cross-layer privacy breaching strategies to human genomic data (using both public genomic databases and other public non-genomic data). We outline the principles and outcomes of each technique, and assess its technological complexity and maturation. We then review potential privacy-preserving countermeasure mechanisms for each threat.

**Keywords:** Genomics, Privacy, Bioinformatics

## 1 Introduction

Today, next-generation sequencing technologies (NGS), are capable of generating a tremendous amount of sequencing data. These technologies allow sequencing the full human genome for as little as few hundred US dollars. As a result, the production of genetic information for research, clinical care, and direct-to-consumer genomics at a rapid pace is no longer impossible from the technological point of view. The availability of human genetic biobanks provides an adequate basis for several important applications and studies. These genetic biobanks involve both genetic data, such as DNA sequence, and health/personal information, such as information about the health, family history, lifestyle, and demographics of an individual. Genomic research typically includes collecting samples from thousands of individuals [2]. Furthermore, a large push is underway to sequence hundreds of thousands to millions of genomes aiming at discovering the functional impact of *de novo* (not inherited from either parent) genetic variations on diseases such as autism and cancer [9].

Accelerating the pace of biomedical breakthroughs and discoveries necessitates not only collecting millions of genetic samples, but also granting open access to the genetic biobanks and databases. According to the Nucleic Acids Research archive [5], this trend has caused the launch of more than one thousand publicly available online genetic databases, in which individuals publicly share their genomic data. Several studies [12, 17, 19] show that the majority (i.e., 69−92%) of the respondents in countries, such as the United States, Japan, and Singapore, have positive attitudes towards genomics research and donating their DNA samples. This willingness of individuals is due to a number of reasons: The most common intention is to support the personalized medicine studies, which involve comparing the genome of patients and healthy individuals. Such studies try to identify the functional impact of certain inherited (or *de novo*) genetic varia-tions

on a disease; aiming at discovering and developing efficient drugs. The second common goal is to learn about their genetic predispositions to diseases and even their genetic compatibilities with potential partners. Last but not least, to identify their distant patrilineal relatives and the potential surnames of their biological fathers.

However, the overwhelming majority of the respondents rank privacy of sensitive information as one of their top concerns. Therefore, proper management of the personal information confidentiality is necessary in order to attain public understanding and support towards genomic research. In addition, transparency of the research aim and proper management of utilization of genetic data should be also maintained in order not to utilize the data beyond the donor's intention. Thus, the biggest challenge of widely utilizing the human genomes and pushing the frontiers of the genetic research is both social and technical. In the literature, there exist reviews addressing genomic privacy (e.g., [4,14]). This paper focuses on the cross-layer attacks against genomic privacy of individuals (using both genomic and non-genomic data) and proposes potential countermeasure mechanisms in a systematic way. We do not cover genome hacking due to poor physical security since it has been extensively discussed in the computer security literature. The rest of the paper is organized as follows. In Section 2, we survey a wide spectrum of known privacy threats to human genomic data. In Section 3, we present our recommendations and guidelines for potential privacy-preserving countermeasure techniques for each threat. Finally, we conclude the paper in Section 4.

## 2 Genetic Privacy Breaching Strategies

In this section, we survey a wide spectrum of privacy threats to human genomic data, as reported by prior research.

### 2.1 Identity tracing by meta-data and side-channel leaks

In such an attack, as illustrated in Fig. 1, the hacker or curious party needs both human genomic data, which is already available online via a certain privacy-preserving mechanism (i.e., hiding the identity information of the owner), and additional metadata, such as basic demographic details and health conditions. These pieces of metadata are exploited to re-identify the owners of the genomes. The metadata can be obtained by a little searching over a number of well-known databases and social networking sites, such as ysearch.org, 23andMe.com, and many others. Such an attack, once it succeeds, can cause serious implications, for instance genetic discrimination, financial loss, and blackmail. A real-life example of this threat was in 1997 when Professor Sweeney [20] successfully identified the medical condition of William Weld, former governor of Massachusetts, using only his demographic data (i.e., date of birth, gender, and 5-digit ZIP code) appearing in the hospital records and voter registration forms that are available to everyone. She also estimated in her study [20] that at most $60-87\%$ of the United States population has unique combinations of date of birth, gender, and ZIP, and hence any data containing all these three attributed is not anonymized. Nonetheless, others have challenged whether Weld was re-identified because he was a public figure or because his demographics were unique [1]. In 2013, Sweeney [21] again showed that it is possible to utilize the demographic data to discover the real identities of the DNA donors even though their names are removed from the published genomic

database. She was able to de-anonymize 241 people from an anonymized public genomic database called Personal Genome Project (PGP). The approach was very similar to her previous attack, besides, in this work, she exploited the side-channel data in the downloaded genomic data files associated with anonymized PGP profiles. Even for some participants, once the downloaded file was uncompressed, the resulting file had a filename that included the actual name of participant.

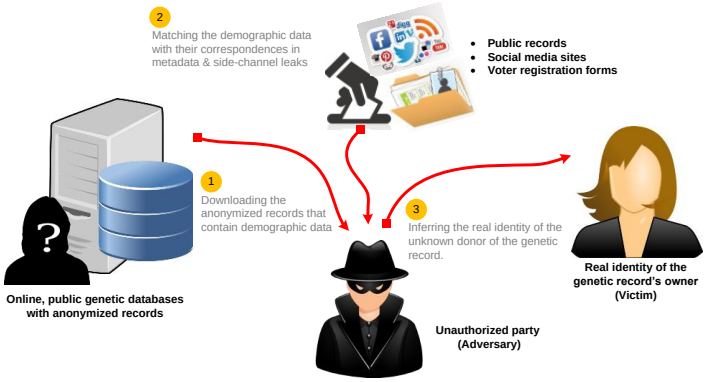


**Fig. 1.** A possible route for identity tracing using both metadata and side-channel leaks.

## 2.2 Identity tracing by genealogical triangulation

In most human societies, surnames are paternally inherited, resulting a correlation with specific Y-chromosome haplotypes. Thus, there are several online public databases (e.g., Ysearch.org and SMGF.org) that collectively contain hundreds of thousands of surname-haplotype records, aiming at helping the public to identify their distant patrilineal relatives and the potential surnames of their biological fathers. However, these services can be exploited by an adversary towards learning the participant's identity, as illustrated in Fig. 2. With the help of surname inferences in addition to the birth year and Zip code, the search results can be narrowed down the identity to few matches that can be investigated individually.

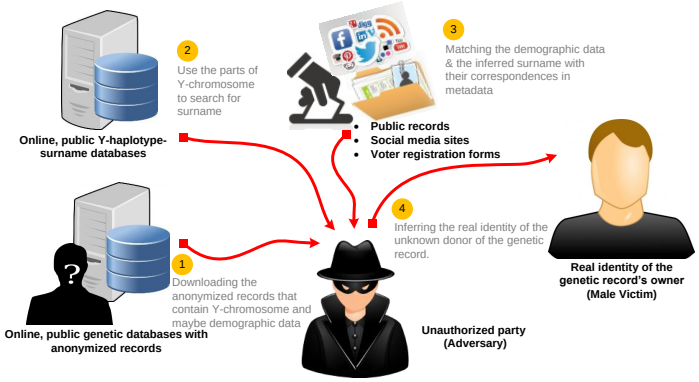


**Fig. 2.** A possible route for identity tracing using genealogical triangulation.

### 2.3 Identity tracing by phenotypic prediction

Visible phenotypes from genetic data could help in identity tracing. Such visible traits with high heritability that can be inferred from DNA include height, eye color, facial morphology, and age [11]. These traits can then be used as quasi-identifiers for decreasing the degree of uncertainty to infer the identity of an individual with the help of public records and social networks as explained in Fig. 3. However, using only these quasi-identifiers for re-identification does not provide high accuracy because of the following limitations. Firstly, a small extent of the phenotypic variability of visible traits are explained in the current genetic studies. The prediction accuracy of the phenotypic variability is not high. Moreover, the population-wide registries of these visible traits are not publicly accessible and searchable. However, rapidly growing social media might help in providing the required data in the future.

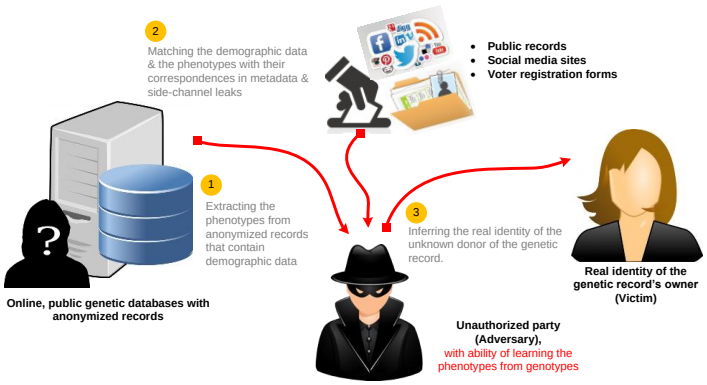


**Fig. 3.** A possible route for identity tracing using phenotypic prediction.

### 2.4 Attribute disclosure attacks via DNA (ADAD)

The main concept of ADAD is when the adversary gains access to the DNA sample of the target. Using the identified DNA, the adversary can search genetic databases with sensitive attributes (e.g., drug abuse) as shown in Fig. 4. Finding the identified DNA in the database reveals the link between the person and the sensitive attribute. Based on [4], three scenarios are identified to illustrate the attribute disclosure attacks: the n=1 scenario, the summary statistic scenario, and the gene expression scenario.

**The n=1 scenario** It is the simplest scenario of ADAD in which the sensitive attribute of the target is associated with the genotype data. By acquiring a small number of autosomal single nucleotide polymorphisms (SNPs) [1], the adversary can simply match the genotype data that is associated with the identity of the individual with the genotype data that is associated with the attribute. Based on [16], a carefully chosen set of 45 SNPs are sufficient to constitute an excellent panel for individual identification. Furthermore, matching a random subset of

[1] SNPs are the main cause for variations in the human genome. They are also responsible for the differences in our phenotypes/traits and genotypes.

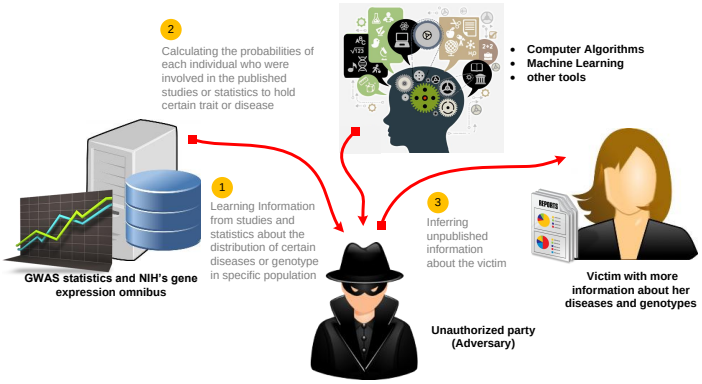


**Fig. 4.** Attribute disclosure attacks via DNA.

300 common SNPs to other data resources could destroy any guarantees of confidentiality by uniquely identify any person [13]. Thus, Genome-Wide Association Studies (GWAS) stores individual genotypes and phenotypes in restricted access area, while the statistics of allele frequencies [2] are stored in the public access area.

**The summary statistic scenario** In spite of the separation, GWAS datasets with allele frequencies of the participants have been exploited by ADAD [6] as follows: The allele frequencies are positively biased towards the target genotypes in the case group compared to the allele frequencies of the general population. Moreover, the analyzed common variations can be exploited to conduct ADAD by integrating the biases in the allele frequencies over a large number of SNPs in GWAS. Therefore, the performance of ADAD is a function of the size of the study and the adversary's prior knowledge.

**The gene expression scenario** Apart from GWAS, the NIH's Gene Expression Omnibus (GEO) databases are also vulnerable to ADAD [18]. The GEO database holds hundreds of thousands of human gene expression profiles and their linked medical attributes. The first step of the algorithm employs a standard expression quantitative trait loci (eQTL) analysis with a reference dataset in order to identify several strong eQTLs and to learn the genotype expression level distributions. Then, it scans the public expression profiles and calculates the probability distributions of the genotypes using a Bayesian approach. Finally, the algorithm matches the target's genotype with the inferred allelic distributions of each expression profile. This technique achieves high accuracy when large-scale simulations are conducted. However, the NIH did not change their policies regarding sharing the human gene expression data due to several limitations and complications of this algorithm.

### 2.5 Completion attacks

In genomics, genotype imputation is a well-studied task in which genetic information can be reconstructed from partial data by completing the missing genotype

[2] The allele frequency represents the incidence of a gene variant at a given gene location in a population gene pool.

values. A well-known example of a completion attack is the inference of Jim Watson's predisposition for Alzheimer's disease from his published genome, despite removing the ApoE locus gene (which is the indicator for Alzheimer's predisposition) from the published data [15]. Completion techniques can be used to predict the genomic information when there is no access to the DNA of a known individual, as shown in Fig. 5. Recently, an individual's profile from OpenSNP.org is used for a completion attack by searching his relatives on Facebook [7]. Eventually, the individual's relatives' genotypes are predicted and their genetic predispositions to Alzheimer's disease are estimated using a Bayesian approach. Moreover, in Iceland, decode genetics succeeded to infer genetic variants of 200,000 living individuals (who never donated their DNA) by using a large reference panel and genealogical information [10]. Consequently, Iceland's Data Protection Authority prohibited the use of this technique until consent is obtained from the individuals.

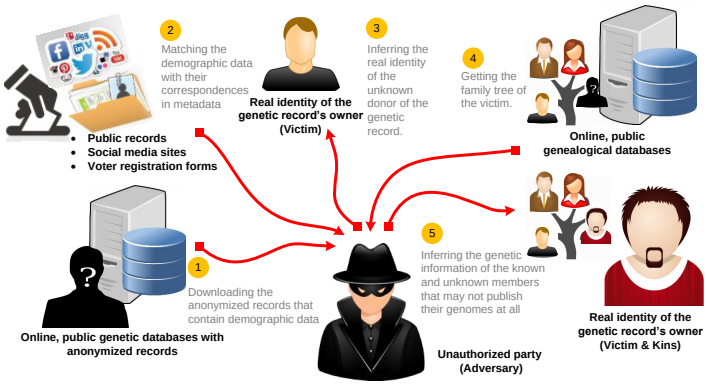


**Fig. 5.** A completion attack.

## 3 Mitigation Techniques

In this section, we survey a wide spectrum of known privacy-preserving techniques against each aforementioned threat and make suggestions to prevent such threats.

### 3.1 Identity tracing by meta-data and side-channel leaks

As discussed in this threat model, metadata can be used for inferring the identities of involved individuals. Hence, any metadata that may decrease the level of privacy, should either be removed from datasets or strictly follow the 2002 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Data covered under HIPAA should follow certain strict formats; dates (e.g. birth, admittance, and discharge dates) would only contain the year, the ZIP code would only have the first 2 digits if the population in the ZIP code is less than 20,000 people, and no explicit identifiers (e.g. name, Social Security numbers, or street addresses) would be present.

### 3.2 Identity tracing by genealogical triangulation

As we discussed, surname can be correlated with the Y-chromosome. The first step towards protecting against this attack depends on the purpose of the genetic

database. If the database provides services for descendants of anonymous sperm donors to identify the surnames of their potential biological father and distant patrilineal relatives, then it should be an access-controlled database. Otherwise, the surname should be removed, or replaced with the given name in haplotype records in order to decrease the ability of connecting surname to unknown's genome. Removing any identifier that can be used for searching in other databases is required to minimize the risk of this kind of attack. Reconstruction attacks based on available online datasets should be performed to measure the connection of surname or other unique identifier with genomic data.

### 3.3 Identity tracing by phenotypic prediction

To prevent this threat, visible traits should be removed if it is unnecessary for datasets and researchers. Also, access control should be considered, because when data is shared publicly, there will be no record of who accessed it, and hence privacy risk will be amplified. Access to sensitive data that acts as quasi-identifiers should be restricted for qualified researchers only by also keeping logs of their access. Nonetheless, predicting a victim's phenotypes is not only based on the revealed information through genetic databases; online social networks can also be a rich source of public sensitive data. Thus, data about visible traits of individuals in public genomic databases as well as other public sources should be restricted (only to qualified researchers or close connections) or removed whenever applicable in order to preserve privacy.

### 3.4 Attribute disclosure attacks via DNA (ADAD)

To address this threat, data perturbation techniques (e.g., differential privacy [3]) can be used for adding noise to the result of a query (on a genomic database) before releasing it publicly. In this way, the reported result will not be much different than original result, but an adversary will not understand if a given individual is in the database or not. Assuming the genomic database includes individuals with a given sensitive attribute, an adversary with prior knowledge can never be sure if that sensitive attribute belong to a specific individual, as similar results will be given when the individual is included in the database or not. However, the added noise should be carefully considered as it will affect the accuracy and the utility of the data at the expense of privacy.

### 3.5 Completion attacks

For this attack that rely on reconstructing genetic information based on partial data, one must consider all available data of each individual that is publicly shared (either by himself, his family members, or genomic researchers) . If with existing completion techniques, one can predict the missing genomic information then specific parts of genomic data should be removed from datasets [8]. Another solution for this attack is using dedicated cryptographic techniques, which enable researchers to access only some parts of the genome by requesting the decryption key from the owner. Such cryptographic solutions can be merged with the reconstruction attack model from [7] to infer the amount of risk that occurs with releasing new portions of data.

## 4 Conclusion

The main concern of publishing the genetic information is the ability to protect the privacy of the sensitive information and its owner. In this work, we surveyed the main five known cross-layer privacy breaching strategies to human genomic data. We outlined the principles and outcomes of each technique, and assessed its technological complexity and maturation. We then gave our guidelines and potential privacy-preserving countermeasure mechanisms for each threat strategy.

## References

1. Barth-Jones, D.C.: The're-identification'of governor william weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (June 4, 2012) (2012)
2. Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S.: A vision for the future of genomics research. Nature 422(6934), 835–847 (2003)
3. Dwork, C.: Differential privacy. In: Encyclopedia of Cryptography and Security, pp. 338–340. Springer (2011)
4. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nature Reviews Genetics 15(6), 409–421 (2014)
5. Galperin, M.Y., Rigden, D.J., et al.: The 2015 nucleic acids research database issue and molecular biology database collection. Nucleic acids research 43(D1), D1–D5 (2015)
6. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., et al.: Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet 4(8), e1000167 (2008)
7. Humbert, M., Ayday, E., Hubaux, J.P., Telenti, A.: Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. pp. 1141–1152. ACM (2013)
8. Humbert, M., Ayday, E., Hubaux, J.P., Telenti, A.: Reconciling utility with privacy in genomics. In: Proceedings of the 13th Workshop on Privacy in the Electronic Society. pp. 11–20. ACM (2014)
9. Iossifov, I., ORoak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., et al.: The contribution of de novo coding mutations to autism spectrum disorder. Nature 515(7526), 216–221 (2014)
10. Kaiser, J.: Agency nixes decode's new data-mining plan. Science 340(6139), 1388–1389 (2013)
11. Kayser, M., de Knijff, P.: Improving human forensics through advances in genetics, genomics and molecular biology. Nature Reviews Genetics 12(3), 179–192 (2011)
12. Kobayashi, E., Sakurada, T., et al.: Public involvement in pharmacogenomics research: a national survey on patients attitudes towards pharmacogenomics research and the willingness to donate dna samples to a dna bank in japan. Cell and tissue banking 12(2), 71–80 (2011)
13. Lin, Z., Owen, A.B., Altman, R.B.: Genomic research and human subject privacy. Science 305(5681), 183 (2004), http://www.sciencemag.org/content/305/5681/183.short
14. Naveed, M., Ayday, E., Clayton, E.W., Fellay, J., Gunter, C.A., Hubaux, J.P., Malin, B.A., Wang, X.: Privacy in the genomic era. To appear in ACM Computing Surveys (2015)
15. Nyholt, D.R., Yu, C.E., Visscher, P.M.: On jim watson's apoe status: genetic information is hard to hide. European Journal of Human Genetics 17(2), 147 (2009)
16. Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C., Furtado, M.R., Kidd, J.R., Kidd, K.K.: Snps for a universal individual identification panel. Human genetics 127(3), 315–324 (2010)
17. Pulley, J.M., Brace, M.M., Bernard, G.R., Masys, D.R.: Attitudes and perceptions of patients towards methods of establishing a dna biobank. Cell and tissue banking 9(1), 55–65 (2008)
18. Schadt, E.E., Woo, S., Hao, K.: Bayesian method to predict individual snp genotypes from gene expression data. Nature genetics 44(5), 603–608 (2012)
19. Storr, C.L., Or, F., Eaton, W.W., Ialongo, N.: Genetic research participation in a young adult community sample. Journal of community genetics 5(4), 363–375 (2014)
20. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)
21. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name. Available at SSRN 2257732 (2013)