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, Bilkent, 06800 Ankara, Turkey
erman@cs.bilkent.edu.tr

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. 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. Genomic research typically includes collecting samples from thousands of individuals, but 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. This trend has caused the launch of more than one thousand publicly available online genetic databases, in which individuals publicly share their genomic data [5]. Several studies [11,16] show that the majority (i.e., 69–92 %) of the respondents have positive attitudes towards genomics research and donating their DNA samples. The most common intention behind it is to support the

J. Garcia-Alfaro et al. (Eds.): DPM and QASA 2015, LNCS 9481, pp. 237–244, 2016.
DOI: 10.1007/978-3-319-29883-2_16

personalized medicine studies. Second, 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. 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., [4,12]). 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. The rest of the paper is organized as follows. In Sect. 2, we survey a wide spectrum of known privacy threats to human genomic data. In Sect. 3, we overview the existing works and present our recommendations and guidelines for potential privacy-preserving countermeasure techniques for each threat. Finally, we conclude the paper in Sect. 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. In general, we assume the existence of a passive attacker who has bounded computational power. In all below threats, the attacker only has access to publicly available genetic databases and other publicly available resources on the Internet.

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 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 Sweeney [17] 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. In 2013, Sweeney [18] 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. 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.

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

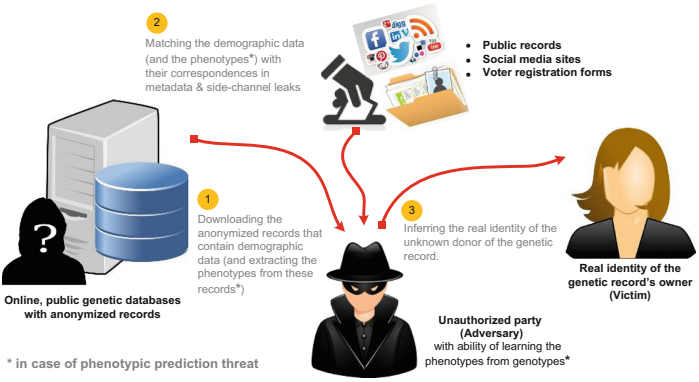

Fig. 1. A possible route for identity tracing using both metadata/side-channel leaks and phenotypic prediction.

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 [6].

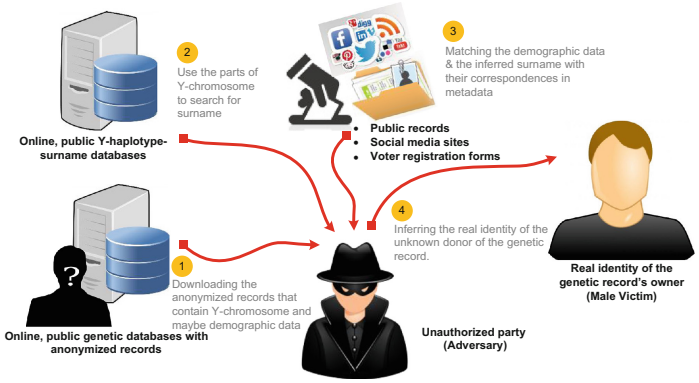

Fig. 2. A possible route for identity tracing using genealogical triangulation.

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 [10]. 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. 1. However, using only these quasi-identifiers for re-identification does not provide high accuracy; as the population-wide registries of these visible traits are not publicly accessible and searchable.

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. 3. 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 is the simplest scenario of ADAD. By acquiring a chosen set of 45 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 [14]. 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. In spite of the separation, GWAS datasets with allele frequencies of the participants have been exploited by the ADAD's summary statistic scenario [7] 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. Apart from GWAS, the NIH's Gene Expression Omnibus (GEO) databases are also vulnerable to the ADAD's gene expression scenario [15]. The GEO database holds hundreds of thousands of human gene expression profiles and their linked medical attributes. However, the NIH did not change their policies regarding sharing the gene expression data due to several complications of this threat.

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 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

[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.

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

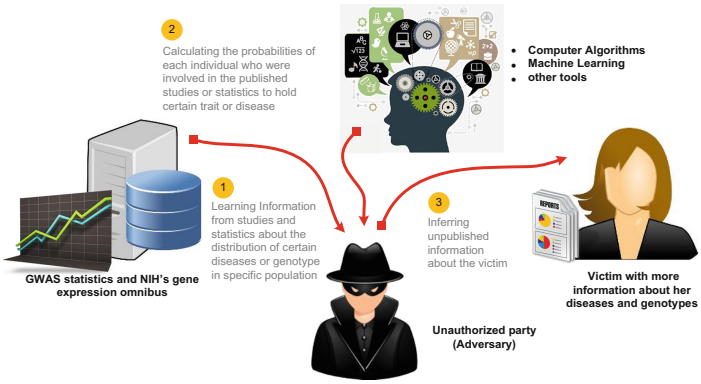

Fig. 3. Attribute disclosure attacks via DNA.

removing the ApoE locus gene (which is the indicator for Alzheimer's predisposition) from the published data [13]. 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. 4.

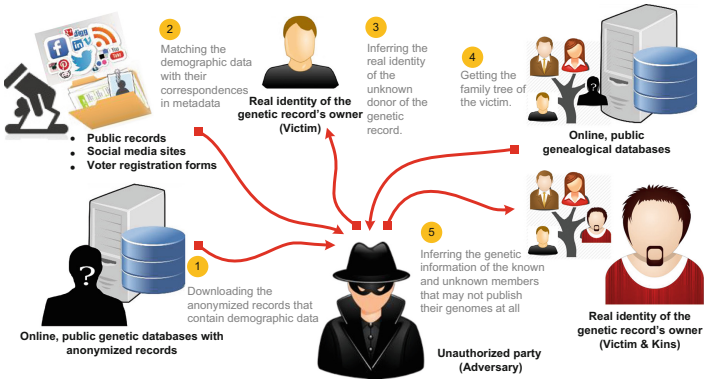

Fig. 4. A possible route for identity de-anonymization using a completion attack.

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. Here, we focus on the scenario, in which genomic data or the results of GWAS are made publicly available. There are also crypto-based mitigation techniques in which genomic data of individuals is stored in a database in encrypted form, and hence it is not publicly available on the Internet. Once other parties (e.g., medical centers) want to do operations on the data, they

apply privacy-preserving techniques and they only obtain the result of the operation without having access to whole data. In this line of research, Ayday et al. proposed privacy-preserving techniques for medical tests and personalized medicine methods [2]. Baldi et al. make use of both medical and cryptographic tools for privacy-preserving paternity tests, personalized medicine, and genetic compatibility tests [3]. Also Ayday et. al developed a technique for privacy-compliant processing of raw genomic data [1]. We note that such scenarios, in which genomic data is not publicly shared, are out-of-the-scope of this paper.

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. Social Security numbers) would be present.

3.2 Identity Tracing by Genealogical Triangulation

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 [6]. 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, 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. 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, and hence privacy risk will be amplified.

3.4 Attribute Disclosure Attacks via DNA (ADAD)

To address this threat, data perturbation techniques (e.g., differential privacy) 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 belongs 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 relies 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. Another solution 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 solutions can be merged with the reconstruction attack model from [8] to infer the amount of risk that occurs with releasing new portions of data.

4 Conclusion

The main concern when publishing anonymized genomic information is usually the privacy of its owner. As it is not trivial to predict the amount of information that will be available to the attacker in today's digital World, existing technical solutions alone are not sufficient to ensure long-term privacy for genomic data donors, and hence their family members. Therefore, there should be a collaborative effort between technical solutions, policies, and legislation (e.g. HIPAA, EU data protection law) to maintain privacy-compliant public genetic databases. As discussed, cryptographic solutions can be an option, but such solutions prevent public availability of genomic data, somehow decreasing the pace of genomic research. This trade-off should also be further investigated.

References

1. Ayday, E., Raisaro, J.L., Hengartner, U., et al.: Privacy-preserving processing of raw genomic data. In: Proceedings of 8th Data Privacy Management (DPM 2013) International Workshop (in conjunction with ESORICS) (2013)
2. Ayday, E., Raisaro, J.L., et al.: Protecting and evaluating genomic privacy in medical tests and personalized medicine. In: Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society, pp. 95–106. ACM (2013)
3. Baldi, P., Baronio, R., et al.: Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 691–702. ACM (2011)

4. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–421 (2014)
5. Galperin, M.Y., et al.: The 2015 nucleic acids research database issue and molecular biology database collection. Nucleic Acids Res. **43**(D1), D1–D5 (2015)
6. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)
7. 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)
8. Humbert, M., Ayday, E., et al.: 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)
9. Iossifov, I., ORoak, B.J., Sanders, S.J., et al.: The contribution of de novo coding mutations to autism spectrum disorder. Nature **515**(7526), 216–221 (2014)
10. Kayser, M., de Knijff, P.: Improving human forensics through advances in genetics, genomics and molecular biology. Nat. Rev. Genet. **12**(3), 179–192 (2011)
11. 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 Tissue Banking **12**(2), 71–80 (2011)
12. 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. ACM Comput. Surv. (CSUR) **48**(1), 6 (2015)
13. Nyholt, D.R., Yu, C.E., Visscher, P.M.: On jim watson's APOE status: genetic information is hard to hide. Eur. J. Hum. Genet. **17**(2), 147 (2009)
14. Pakstis, A.J., Speed, W.C., Fang, R., Hyland, F.C., et al.: SNPs for a universal individual identification panel. Hum. Genet. **127**(3), 315–324 (2010)
15. Schadt, E.E., Woo, S., Hao, K.: Bayesian method to predict individual SNP genotypes from gene expression data. Nat. Genet. **44**(5), 603–608 (2012)
16. Storr, C.L., Or, F., et al.: Genetic research participation in a young adult community sample. J. Commun. Genet. **5**(4), 363–375 (2014)
17. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(05), 557–570 (2002)
18. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name. Available at SSRN 2257732 (2013)