Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200158

GenoShare: Supporting Privacy-Informed Decisions for Sharing Individual-Level Genetic Data

Jean Louis RAISARO [a,1], Juan Ramón TRONCOSO-PASTORIZA [b], Yamane EL-ZEIN [c], Mathias HUMBERT [d], Carmela TRONCOSO [b], Jacques FELLAY [a] and Jean-Pierre HUBAUX [b]

[a] *Lausanne University Hospital, Lausanne, Switzerland*
[b] *EPFL, Lausanne, Switzerland*
[c] *CSEM, Neuchâtel, Switzerland*
[d] *Armasuisse, Lausanne, Switzerland*

Abstract. One major obstacle to developing precision medicine to its full potential is the privacy concerns related to genomic-data sharing. Even though the academic community has proposed many solutions to protect genomic privacy, these so far have not been adopted in practice, mainly due to their impact on the data utility. We introduce GenoShare, a framework that enables individual citizens to understand and quantify the risks of revealing genome-related privacy-sensitive attributes (e.g., health status, kinship, physical traits) from sharing their genomic data with (potentially untrusted) third parties. GenoShare enables informed decision-making about sharing exact genomic data, by jointly simulating genome-based inference attacks and quantifying the risk stemming from a potential data disclosure.

Keywords. genomic privacy, privacy-conscious tools, risk quantification, inference

1. Introduction

The growing number of health data breaches [1] and the recent revelations about the use of free genealogical databases for law enforcement purposes [2] are increasingly eroding public trust in social media, genealogy websites (e.g., 23andMe and Ancestry.com) and personalized medicine initiatives (e.g., the U.S. All of Us Research Program and the U.K. 100,000 Genomes Project), worldwide. Guaranteeing individuals' genetic privacy is becoming a major challenge for those who want to encourage sharing of medical data for research purposes and better personalized medicine. In response to this demand, the computer security community has made a remarkable effort in providing solutions to secure the storage [3], and the processing of genetic data [4,5,6,7,8]. Yet, the issue of trust goes beyond the provision of secure storage and processing. It is also intimately related to the question of *control* over what data is shared and what information it conveys. In

[1] Corresponding Author: Jean Louis Raisaro, Lausanne University Hospital CH-1010, Lausanne, Switzerlands; E-mail: Jean.Raisaro@chuv.ch.

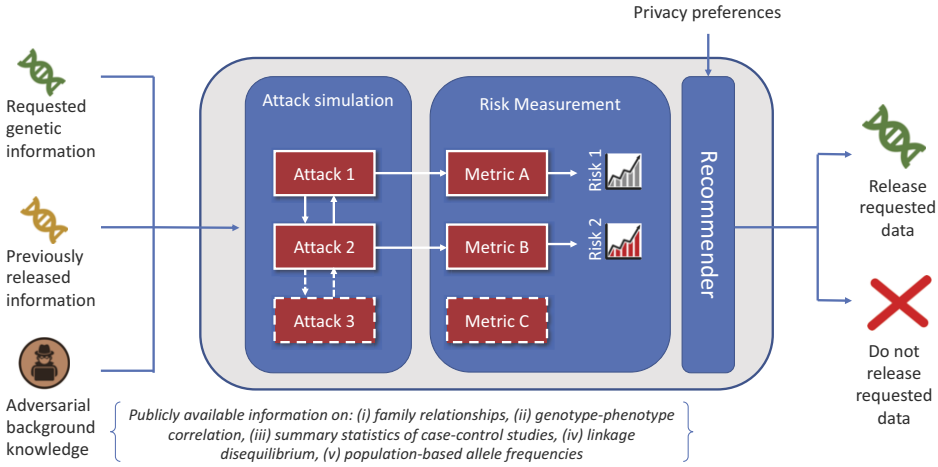

Figure 1. High-level representation of the GenoShare framework.

this paper, we introduce *GenoShare*, the first decision-support framework that provides individuals with guidance at the time of deciding what data can be shared without risking the undesired disclosure of sensitive information.

2. Methods

Let's assume an individual wishes to share part of her genetic profile with a public or private third party in order to join a personalized medicine study or obtain some genetic-related services, but she is concerned about her genetic privacy. Upon reception of a request for genetic data sharing, such an individual can use GenoShare to quantify the risk of sensitive attribute disclosure associated to revealing those data. To this end, GenoShare simulates the combination of inference attacks relevant to the privacy concerns of the individual (e.g., revealing the participation in a sensitive study, revealing undisclosed kinship, revealing predisposition to certain health conditions) and, based on the information available to the adversary, measures the risk of a potential privacy breach. We consider that a realistic adversary could have access to: (i) part of the individual's genetic profile obtained in the past, (ii) the statistical regularities within the genetic population to which the individual belongs (e.g., linkage disequilibrium, allele frequencies), (iii) the effect-size statistics of the genotype-phenotype association, (iv) the genetic profiles of the individual's relatives who might have already joined the study, and the (v) aggregate statistics from other studies in which the individual might have participated. GenoShare is a modular framework and enables different parameterizations (i.e., different levels of adversarial prior knowledge) and the instantiations of different inference attacks, according to the data-sharing context (Fig. 1). Furthermore, GenoShare provides a mechanism, based on the generation of synthetic genomes called *avatars*, to protect individuals' genetic privacy from inferences stemming from decision to not release the requested data. Details about the algorithms underpinning GenoShare are described in the extended version of this paper [9].

3. Results

To show how GenoShare can, in practice, support privacy-conscious decisions when sharing genetic data, we implement an operational prototype in Python with a Web user interface. We instantiate it with three of the most important genomics-oriented inference attacks that we re-adapted to work on a joint manner and on adversarial partial knowledge: the phenotype inference attack, the membership inference attack and the kinship inference attack. For our experiments, we used individual genetic profiles from the 1,000 Genomes Project [10] and considered different data-sharing use cases where the adversary has access to an increasing amount of information about the targeted individual. For example, for a given individual, GenoShare shows that revealing her genetic variants related to schizophrenia (400 variants) introduces a risk of almost 100% of leaking the value of her predisposition to bipolar disorder and her participation in a study with less than 50 people, and a risk of 60% of leaking her kinship with a first-degree relative who might be known by the adversary. Details and figures about the experiments run to validate GenoShare are described in the extended version of this paper [9].

4. Discussion

Academic solutions for privacy-preserving sharing of genetic data that provide formal guarantees of privacy, such as differential privacy, have mostly focused on data perturbation (e.g., differential privacy). Such solutions, however, have not been adopted by the medical sector as (i) they are only suitable for protecting genetic privacy when aggregate information is shared and (ii) they damage the utility of the data such that is then unusable by practitioners. As opposed to these solutions, GenoShare quantifies the risk of sensitive attribute disclosure when *individual and exact* genetic data is released by using novel meaningful sensitive attribute-oriented metrics and by considering realistic adversaries with limited and parametrizeable background knowledge. Finally, to the best of our knowledge, GenoShare is also the first framework to jointly consider relevant attacks in genomic privacy in presence of incomplete information. Therefore, it provides a principled answer to the privacy concerns that affecting the genomic community, and thus it is a firm step forward to enable the responsible and privacy-respecting use of genomic data in research and medical environments.

References

[1] U.S. Department of Health and Human Services . Breach portal: Notice to the secretary of hhs breach of unsecured protected health information. `https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf`. Last Accessed: October 30, 2019.

[2] Natalie Ram, Christi J Guerrini, and Amy L McGuire. Genealogy databases and the future of criminal investigation. *Science*, 360(6393):1078–1079, 2018.

[3] Zhicong Huang, Erman Ayday, Jacques Fellay, Jean-Pierre Hubaux, and Ari Juels. Genoguard: Protecting genomic data against brute-force attacks. In *IEEE Symposium on Security and Privacy*, pages 447–462. IEEE Computer Society, 2015.

[4] Pierre Baldi, Roberta Baronio, Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. In *ACM Conference on Computer and Communications Security, (CCS)*, pages 691–702, 2011.

[5] George Danezis and Emiliano De Cristofaro. Fast and private genomic testing for disease susceptibility. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014.

[6] Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K. Reiter, and Zheng Dong. Privacy-preserving genomic computation through program specialization. In *ACM Conference on Computer and Communications Security, (CCS)*, pages 338–347, 2009.

[7] Paul J McLaren, Jean Louis Raisaro, Manel Aouri, Margalida Rotger, Erman Ayday, István Bartha, Maria B Delgado, Yannick Vallet, Huldrych F Günthard, Matthias Cavassini, et al. Privacy-preserving genomic testing in the clinic: a model using HIV treatment. *Official journal of the American College of Medical Genetics and Genomics*, 2016.

[8] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):6, 2015.

[9] Jean Louis Raisaro, Carmela Troncoso, Mathias Humbert, Zoltan Kutalik, Amalio Telenti, and Jean-Pierre Hubaux. Genoshare: Supporting privacy-informed decisions for sharing exact genomic data (extended version). Technical report, EPFL infoscience, 2018.

[10] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.