Privacy and Security in the Genomic Era

Erman Ayday
Bilkent University
Computer Engineering Department
Ankara, Turkey
erman@cs.bilkent.edu.tr

Jean-Pierre Hubaux
EPFL
School of Computer and Communications
Lausanne, Switzerland
jean-pierre.hubaux@epfl.ch

ABSTRACT

With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, the research in genomics has gained speed in paving the way to personalized (genomic) medicine, and geneticists need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their (genetic) predispositions to diseases, their ancestries, and even their (genetic) compatibilities with potential partners. This trend has also caused the launch of health-related websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP or 23andMe). On the other hand, genomic data carries much sensitive information about its owner. By analyzing the DNA of an individual, it is now possible to learn about his disease predispositions (e.g., for Alzheimer's or Parkinson's), ancestries, and physical attributes. The threat to genomic privacy is magnified by the fact that a person's genome is correlated to his family members' genomes, thus leading to interdependent privacy risks. This *short tutorial* will help computer scientists better understand the privacy and security challenges in today's genomic era. We will first highlight the significance of genomic data and the threats for genomic privacy. Then, we will present the high level descriptions of the proposed solutions to protect the privacy of genomic data and we will discuss future research directions. No prerequisite knowledge on biology or genomics is required for the attendees of this proposal. We only require the attendees to have a slight background on cryptography and statistics.

1. OVERVIEW OF THE TUTORIAL

We briefly provide an overview of the proposed tutorial in the following.

1.1 What is Genomic Data?

To familiarize computer scientists with the topic, we will first briefly provide a gentle background on genomic data. Without going into details, we will describe the pipeline for the generation of digital genomic data starting from the biological samples and going all the way to the utilization of genomic data for different purposes.

1.2 Why is Genomic Data Special?

Genomic data has many unique features. To motivate the discussion, we will highlight some of these features. One notable feature is how it is static and of long-lived value. Most medical data, such as body temperature and blood pressure, are of relatively short term value, whereas genomic data changes little over a lifetime and may have value that lasts for decades.

While DNA has been used for some time in parentage tests, it can be generalized from such studies to enable broader inference of kinship relations. Services such as Ancestry.com and 23andme.com already offer kinship services based on DNA testing. While a substantial portion of an individual's DNA is in common with that of her relatives, it is also unique to her (unless she has an identical twin). This has another set of implications about potential use of genomic data, like its ability to link to her personally, a property that makes DNA testing useful in criminal forensics.

Another essential value of DNA relates to its ability for diagnosing problems in health and behavior. Tests are able to demonstrate increased likelihood for conditions like macular degeneration in old age and Alzheimer's (the most common form of dementia). Although these are often probabilities, they can have diagnostic value as well as privacy ramifications. This power for good and bad has led genomic data to have a certain "mystique", which has been promoted by scientists and the media.

1.3 Privacy and Security Related Implications of Genomic Data

Genomic data already has many uses in several areas in real-life. In general, such areas can be summarized as (i) healthcare, (ii) research, (iii) direct-to-consumer (DTC) genomics, and (iv) forensics. Such wide usage of genomic data also raises serious privacy-related concerns on the data. Thus, we will discuss some well-known and important attacks on privacy of genomic data.

It has been shown that anonymization is an ineffective technique for sharing genomic data [9, 10]. For instance, genomic variants on the Y chromosome are correlated with the last name (for males) and this last name can be inferred using public genealogy databases. Also, unique features in patient-location visit patterns in a distributed health care environment can be used to link the genomic data to the identity of the individuals in publicly available records [18].

CCS'16 October 24-28, 2016, Vienna, Austria

ACM ISBN 978-1-4503-4139-4/16/10.

DOI: http://dx.doi.org/10.1145/2976749.2976751

The identity of a participant of a genomic study can also be revealed by using a second sample, that is, part of the DNA information from the individual and the results of the corresponding clinical study [11, 23, 9]. Homer et al. prove that the presence of an individual in a case group can be determined by using aggregate allele frequencies and his DNA profile [11]. This work compelled the NIH, as well as the Wellcome Trust in the UK, to remove all publicly available aggregate genomic data from their websites. Wang et al. showed a higher risk that individuals can actually be identified from a relatively small set of statistics, such as those routinely published in GWAS papers [23]. Humbert et al. evaluated the genomic privacy of an individual threatened by his relatives revealing their genomes [13]. Very recently, Shringarpure and Bustamante showed that it is possible to identify members of a beacon server[1] by using the results of the queries [20].

1.4 Proposed Solutions for Privacy and Security of Genomic Data

Next, focusing on the below issues, we will highlight some of the existing solutions for the security and privacy of genomic data.

Read mapping and alignment: Jha et al. proposed a privacy-preserving implementation of fundamental genomic computations for sequence alignment [17]. Chen et al. proposed a privacy-preserving method to align short sequences by outsourcing the computation to the cloud [6].

Processing or raw genomic data: Ayday et al. proposed a technique for privacy preserving storage, retrieval, and processing of raw genomic data [1].

Storage: Traditional cryptographic techniques are prone to brute-force attacks. It is not too far-fetched to imagine that a third-party in possession of an encrypted genome might be able to decrypt it years or decades later. To counter this problem, Huang et al. proposed a scheme named GenoGuard to protect genomic data against brute force attacks [12].

Clinical use of genomic data: De Cristofaro et al. proposed a secure protocol between two parties for testing genomic sequences without the leaking of any private information about the genomic sequence or the nature of the test [7]. Ayday et al. proposed a scheme to protect the privacy of users' genomic data while enabling medical units to access the genomic data in order to conduct medical tests or to develop personalized medicine methods [2, 3].

Genomic research: Efforts to provide privacy-preserving use of genomic data in research can be put in three main categories: (i) techniques based on differential privacy, in which a controlled noise is added to the result of a query (to a genomic database) [15, 8, 21], (ii) techniques based on cryptography, in which the use of homomorphic encryption, secure hardware, or secure multiparty computation are proposed for privacy-preserving genomic research [16, 5], and (iii) techniques based on optimization, in which the goal is to maximize the amount of publicly shared genomic data while also complying the privacy preferences of individuals [14].

Comparing genomes: Troncoso-Pastoriza et al. proposed an algorithm for private string searching on the DNA sequence by using a finite state machine [22]. Baldi et al. made use of private-set intersection and present an effective algorithm for privacy-preserving substring matching on DNA sequences [4]. Naveed et al. proposed using functional encryption for privacy-preserving similarity test on genomic data [19]. Recently, Wang et al. proposed private edit distance protocols to find similar patients (e.g., across several hospitals) [23].

1.5 Open Research Directions

Finally, we will discuss the open research problems about the security and privacy of genomic data. In particular, we will discuss (i) genomic data sharing between different entities, (ii) privacy vs. utility of genomic data, (iii) integrating genomic data in individuals' electronic health records, (iv) access control issues, (v) credibility and liability issues of genomic data, and (vi) standardization efforts by Global Alliance for Genomics and Health (GA4GH).

Personalized medicine brings the promise of better diagnoses, better treatments, a higher quality of life, and increased longevity. To achieve these noble goals, it exploits a number of revolutionary technologies, including genome sequencing and DNA editing, as well as wearable devices and implantable or even edible biosensors. In parallel, the popularity of "quantified self" gadgets shows the willingness of citizens to be more proactive with respect to their own health. Yet, this evolution opens the door to all kinds of abuses, notably in terms of discrimination, blackmailing, stalking, and subversion of devices. We are convinced that this tutorial will significantly contribute to the awareness of our community with respect to the magnitude of the challenge and will help it design timely and effective solutions.

2. SPEAKER BIOGRAPHIES

Erman Ayday is an assistant professor of computer science at Bilkent University, Ankara, Turkey. Before that he was a post-doctoral Researcher at EPFL, Switzerland, working with Prof. Jean-Pierre Hubaux. He received his M.S. and Ph.D. degrees from Georgia Tech in 2007 and 2011, respectively. Erman's research interests include privacy-enhancing technologies (including big data and genomic privacy), wireless network security, trust and reputation management, and applied cryptography. Erman is the recipient of Distinguished Student Paper Award at IEEE S&P 2015 and 2011 ECE GRA Excellence Award from Georgia Tech. Other various accomplishments of Erman include several patents, research grants, and H2020 Marie Curie fellowship.

Jean-Pierre Hubaux is a full professor at the School of Information and Communication Sciences of EPFL. Through his research, he contributes to laying the foundations and developing the tools to protect privacy in tomorrow's hyperconnected world. He is focusing notably on network privacy and security, with an emphasis on mobile/wireless networks and on data protection, with an emphasis on health-related data and especially genomic data. He has worked on the topic of genome privacy since 2011 and has designed cryptographic solutions, notably in collaboration with the Lausanne University Hospital (CHUV) and the EPFL School of Life Sciences. He has co-chaired the first workshop devoted to the topic (in Dagstuhl, Germany, in 2013) and is a co-founder and chair of the steering committee of the International Workshop on Genome Privacy and Security.

[1]Beacon hosts genomic data of people with a certain disease. Researchers can query the beacon by asking for the existence of an entry, that is a certain allele at a specific position. The beacons response can be either "yes" or "no".

3. REFERENCES

[1] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux. Privacy-preserving processing of raw genomic data. *DPM*, 2013.

[2] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J.-P. Hubaux. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. *Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech)*, 2013.

[3] E. Ayday, J. L. Raisaro, J. Rougemont, and J.-P. Hubaux. Protecting and evaluating genomic privacy in medical tests and personalized medicine. *WPES*, 2013.

[4] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. *Proceedings of ACM CCS '11*, pages 691–702, 2011.

[5] M. Canim, M. Kantarcioglu, and B. Malin. Secure management of biomedical data with cryptographic hardware. *IEEE Transactions on Information Technology in Biomedicine*, 16(1), 2012.

[6] Y. Chen, B. Peng, X. Wang, and H. Tang. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *NDSS'12: Proceeding of the 19th Network and Distributed System Security Symposium*, 2012.

[7] E. De Cristofaro, S. Faber, and G. Tsudik. Secure genomic testing with size- and position-hiding private substring matching. *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, pages 107–118, 2013.

[8] S. E. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. *Proceedings of the IEEE ICDMW '11*, Dec. 2011.

[9] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.*, 84:251–258, 2009.

[10] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science: 339 (6117)*, January 2013.

[11] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4, Aug. 2008.

[12] Z. Huang, E. Ayday, J.-P. Hubaux, J. Fellay, and A. Juels. Genoguard: Protecting genomic data against brute-force attacks. *n Proceedings of IEEE Symposium on Security and Privacy*, 2015.

[13] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Addressing the concerns of the Lacks family: Quantification of kin genomic privacy. *CCS*, 2013.

[14] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. Reconciling utility with privacy in genomics. *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, 2014.

[15] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. *KDD*, pages 1079–1087, 2013.

[16] M. Kantarcioglu, Wei Jiang, Ying Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5):606–617, 2008.

[17] S. Jha L. Kruger and V. Shmatikov. Towards practical privacy for genomic computation. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 216–230, 2008.

[18] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: Using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37, Jun. 2004.

[19] M. Naveed, S. Agrawal, M. Prabhakaran, X. Wang, E. Ayday, J.-P. Hubaux, and C. Gunter. Controlled functional encryption. *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014.

[20] S.S. Shringarpure and C.D. Bustamante. Privacy risks from genomic data-sharing beacons. *Am J Hum Genet*, 97(5):631–646, February 2015.

[21] F. Tramèr, Z. Huang, J.-P. Hubaux, and E. Ayday. Differential privacy with bounded priors: Reconciling utility and privacy in genome-wide association studies. *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1286–1297, 2015.

[22] J.R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik. Privacy preserving error resilient DNA searching through oblivious automata. *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, 2007.

[23] R. Wang, X. Wang, Z. Li, H. Tang, M. K. Reiter, and Z. Dong. Privacy-preserving genomic computation through program specialization. *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pages 338–347, 2009.