When Biology Gets Personal: Hidden Challenges of Privacy and Ethics in Biological Big Data

Gamze Gürsoy*

Computational Biology and Bioinformatics Program, Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, 06511, USA
Email: gamze.gursoy@yale.edu

Arif Harmanci

Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, 77030, USA
Email: arif.o.harmanci@uth.tmc.edu

Haixu Tang†

School of Informatics, Computing and Engineering, Indiana University Bloomington, Bloomington, IN, 47405, USA
Email: hatang@indiana.edu

Erman Ayday

Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, 44106, USA
Email: exa208@case.edu

Steven E. Brenner#

University of California Berkeley, CA, 94720-3012, USA
Email: brenner@compbio.berkeley.edu

High-throughput technologies for biological data acquisition are advancing at an increasing pace. Most prominently, the decreasing cost of DNA sequencing has led to an exponential growth of sequence information, including individual human genomes. This session of the 2019 Pacific Symposium on Biocomputing presents the distinctive privacy and ethical challenges related to the generation, storage, processing, study, and sharing of individuals' biological data generated by multitude of technologies including but not limited to genomics, proteomics, metagenomics, bioimaging, biosensors, and personal health trackers. The mission is to bring together computational biologists, experimental biologists, computer scientists, ethicists, and policy and lawmakers to share ideas, discuss the challenges related to biological data and privacy.

Keywords: biological data privacy, genomics, genetic testing

* This work is partially supported by NIH grant U01EB023686.

† This work is partially supported by NIH grant U01EB023685 and NSF grant CNS-1408874.

This work is partially supported by NIH grant U01EB023686 and NIH grant U41HG007346.

1. Introduction

Data privacy is an important topic of debate crossing many different fields such as ethics, sociology, law, political science and forensic science. Thanks to the rapid reduction of the DNA sequencing cost in the past decade, the number and the volume of available genomic data have exponentially increased [1]. Hence, individuals' genomic data has recently emerged as one of the major foci of studies on privacy as availability of genetic information gives rise to privacy concerns [2]. For example, individuals express concern that genetic predisposition to diseases may bias insurance companies or enable unlawful discrimination by employers [3,4,5]. On a larger scale, imagine the economic repercussions had it been leaked that the CEO of Apple Computer had pancreatic cancer and was not adhering to a typical oncological regimen. Recently it has been also shown that high throughput molecular phenotype datasets such as functional genomic and metabolomics measurements, and microbiome measurements increased the number of quasi-identifiers for participating individuals that can be used by adversaries for re-identification purposes [6,7,8,9]. In addition, the emergence of electronic health records (EHR) with the rise of personalized medicine makes patients vulnerable to breaching privacy. These results indicate that privacy concerns over sharing personalized biological data will increase quickly with the increase in the number of genetic and ancestry testing companies, which collect and distribute very large amount of health related data, including genetic data (such as 23andMe) or health and fitness tracking data (such as fitbit). The data collection and sharing methods that these companies use call for a public discussion of privacy considerations around these new concepts. Moreover, the recent arrest of the Golden State Killer, through long-range familial search on consumer genomics databases, sparked questions over the risk of re-identification based on genetic testing taken by a relative. Recent two studies showed the statistical risk of identifying relatives as being high by using long-range familial searches [10,11].

Protecting the privacy of study participants has emerged as an important issue in genotype-phenotype association studies. Several studies investigated whether a genome of an individual can be detected in a mixture [12,13,14,15]. As a result, various counter-measures have been proposed to protect participant privacy [16]. As the number of genotype-phenotype datasets increase, new routes for breaching privacy such as cross-referencing multiple databases opened up [17,18]. Access control, data anonymization and cryptographic techniques were studied to prevent privacy breaches [4]. Ultimately, the ability to keep these data private is unclear, and so preparations for both small and catastrophic leaks must be made [5]. As the technologies increase, new data types are being released and more studies to investigate the potential privacy breaches will be needed. This area of research has become more and more interdisciplinary, where ethics researchers inform researchers who work on privacy-preserving techniques, while these techniques inform policymakers to reform laws and policies.

On the other side of the privacy problem, the benefit and importance of open data sharing is widely acknowledged, as the solutions such as access control or cryptographic techniques delay the access to the data by average researchers either by creating bureaucratic bottlenecks or technical challenges. Open data sharing harbors the collaboration between different biomedical researchers by allowing rapid exchange of the information. Funding agencies and research organizations are increasingly supporting new means of data sharing and new requirements for making data publicly available while preserving participants' privacy [19]. This increases the value of the techniques and policies that prevent the sensitive information leakage while promoting data sharing.

The papers featured in this session represent various aspects of biological data privacy highlighting a number of problems and solutions that need to be addressed to protect privacy of individuals while encouraging open data sharing. Topics in this session include making inferences on complex phenotypes in

large biobanks, patient re-identification through electronic health records and countermeasures, privacy-preserving GWAS studies as well as efforts on improving informed consents for AllofUs research project.

2. Podium Presentations

After the seminal work by Homer et. al [12], the policies on how to share GWAS results have been changed and only summary statistics are allowed to share publicly. **Gasdaska et al. [20]** explore the possibility of using these summary statistics to make inferences about the hidden, complex phenotypes that are derived from two or more phenotypes. This potentially reveals information about the participants that they may not want to disclose. Investigators validated their statistical derivations on simulated and real datasets.

As **A. Gasdaska** and colleagues [20] showed that sharing statistical aggregates from GWAS might have sensitive information leakages and also demonstrated that how complex phenotypes can be analyzed in terms of simple phenotypes in a privacy preserving fashion, **S. Simmons** and colleagues [21] showed us how we could reduce this leakage by introducing a Laplacian noise to the released data. The investigators presented a novel method for measuring privacy loss in GWAS summary statistics. This was achieved by providing a probabilistic formulation for measuring the risk of releasing summary statistics as the posterior probability of an individual being in the cohort. With the introduction of an MCMC-based method for computing this posterior probability, the authors reduced the degree of privacy leakage with the same amount of data released. This work presented interesting ideas on how to control the privacy risk by setting a noise level and the amount of data to be released.

K. Johnson and colleagues [22] studied the privacy leakages of Electronic Health Records. They showed that lab tests can be used as quasi-identifiers for patients for re-identification of patients' medical records. The investigators used the EHR at Mount Sinai Hospital as a case study. This study took an even more interesting turn when they used variational auto-encoder to encode the lab test results to reduce the privacy risk of re-identification. They showed a substantial decrease in re-identification risks when the lab tests were stored as latent variables while the encoded test results still provide almost the same utility as original results when compared in terms of classification accuracy. Although further work is required to show how decoding-encoding will be achieved in this new representation, the novel idea of storing data will open up the doors for storing other kind of private data in the future.

3. Posters with Published Papers

This year's poster session with papers published in the proceedings will feature a unique study that has not been explored at PSB before by **M. Doerr** and colleagues [23]. They designed a study to give a comprehensive overview of existing jurisdictions for the informed consent process for the AllofUs initiative and its compliance with the state/territory regulations. This study will be of great interest for the investigators of the AllofUs project, which aims to collect a vast amount of biomedical data from a million of Americans.

4. Acknowledgments

We would like to thank the members of the program committee for reviewing all submissions and providing expert critiques used in evaluating manuscripts for inclusion into the session and the PSB proceedings. We would also like to thank the PSB 2018 chairs and Tiffany Murray of Stanford University for their efforts in organizing the meeting.

References

1. Sboner A, Mu X, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biology, 2011;12(8):125.
2. Brenner SE. Be prepared for the big genome leak. Nature, 2013;498:139
3. Joly Y, Dyke SOM, Knoppers BM, Pastinen T. Are Data Sharing and Privacy Protection Mutually Exclusive? Cell, 2016;167(5):1150-1154.
4. Joly Y, Feze IN, Song L, Knoppers BM. Comparative Approaches to Genetic Discrimination: Chasing Shadows? Trends Genet, 2017;33(5):299-302.
5. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat. Rev. Genet., 2014;15(6):409-421.
6. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nature Methods, 2016;13(3):251-256.
7. Harmanci A, Gerstein M. Analysis of Sensitive Information Leakage in Functional Genomics Signal Profiles through Genomic Deletions. Nature Communications, 2018; 9 (1), 2453
8. Gürsoy G, Harmanci A, Green M, Navarro F, Gerstein M. Sensitive information leakage from functional genomics data: Theoretical quantifications & practical file formats for privacy preservation, 2018, Biorxiv
9. Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJ, Huttenhower C. Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci U S A.* 112(22):E2930-8 (2015).
10. Erlich, Y. *et al.* Identity inference of genomic data using long-range familial searches. *Science.* (2018).
11. Kim J, Edge MD, Algee-Hewitt BFB, Li JZ, Rosenberg NA. Statistical Detection of Relatives Typed with Disjoint Forensic and Biomedical Loci. *Cell.* (2018).
12. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4, e1000167 (2008).
13. Im, H.K., Gamazon, E.R., Nicolae, D.L. & Cox, N.J. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* 90, 591–598 (2012).
14. Lunshof, J.E., Chadwick, R., Vorhaus, D.B. & Church, G.M. From genetic privacy to open consent. *Nat. Rev. Genet.* 9, 406–411 (2008).
15. Church, G. *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet.* 5, e1000665 (2009).
16. Jiang X, Zhao Y, Wang X, Malin B, Wang S, Ohno-Machado L, Tang H. A community assessment of privacy preserving techniques for human genomes. BMC Med Inform Decis Mak. 2014;14 Suppl 1:S1.
17. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science, 2013;339(6117):321-324
18. Sweeney L. Simple demographics often identify people uniquely. Carnegie Mellon University, unpublished, 2000.
19. National Institute of Health data sharing policy. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-110.html

20. Gasdaska A, Friend D, Chen R, Westra J, Zawistowski M, Lindsey W, Tintle N. Leveraging summary statistics to make inferences about complex phenotypes in large biobanks.
21. Simmons S, Berger B, Sahinalp C. Protecting Genomic Data Privacy with Probabilistic Modeling
22. Johnson KW, De Freitas JK, Glicksberg BS, Bobe JR, Dudley JT. Evaluation of patient re-identification using laboratory test orders and mitigation via latent space variable.
23. Doerr M, Grayson S, Moore S, Suver C, Wilbanks J, Wagner J. Implementing a universal informed consent process for the *All of Us* Research Program