Computational Frameworks for Genomic Privacy Analysis

# BACKGROUND

Privacy is one of the most important topics of debate in data science that stands at the corner of many different fields, including ethics, sociology, law, political science, and forensic science. Recently, genomics has emerged as one of the major foci of studies on privacy. This can mainly be attributed to the advancement of technologies for high throughput biomedical data acquisition that bring about a surge of datasets1,2. Among these, high throughput molecular phenotype datasets, like functional genomic and metabolomic measurements, substantially grow the list of the *quasi-identifiers* (such as birth date, ZIP code, gender3) for participating individuals, which can be used by an adversary for re-identification of the identities. With the recent announcement of Precision Medicine Initiative4, a large body of datasets are to be generated and shared among researchers5. The National Institutes of Health also released the plans to encourage public access to biomedical datasets from scientific studies 5–7. Considering the fact that one does not need many identifiers to uniquely pinpoint an individual3,8,9, these datasets have the potential to exacerbate the risk of privacy breach.

Many consortia, like GTex10, ENCODE11, 1000 Genomes12, and TCGA13, are generating large amount of personalized biomedical datasets. Coupled with the generated data, sophisticated analysis methods are being developed to discover correlations between genotypes and phenotypes, some of which can contain sensitive information like disease status. Although these correlations are useful for discovering how genotypes and phenotypes interact, they could also be utilized by an adversary in a linking attack for matching the entries in genotype and phenotype datasets. For example, when a phenotype dataset is available, the adversary can utilize the genotype-phenotype correlations to statistically predict the genotypes, compare the predicted genotypes with the entries in another dataset that contains genotypes. For the entries that are correctly matching, he/she can reveal sensitive phenotypes of the individuals and characterize them. Even when the strength of each genotype-phenotype correlation is not high, the availability of a large number of genotype-phenotype correlations increases the scale of linking. In fact, an adversary can perform correct linking with relatively small number of genotypes14,15.

Different aspects of privacy have been intensely studied. Recently, genomic privacy is receiving much attention as a result of the deluge of personalized genomics datasets that are being generated16,17. With the increase in the number of large scale genotyping and phenotyping studies, the protection of privacy of participating individuals emerged as an important issue. Homer et al18 proposed a statistical testing procedure that enables testing whether a genotyped individual is in a pool of samples, for which only the allele frequencies are known. Im et al19 showed that, given the genotypes of a large set of markers for an individual, an attacker can reliably predict whether the individual participated to a QTL study or not. These attacks, which we refer to as “detection of a genome in a mixture”, are one type of attacks on privacy. There is yet another important attack where the attacker links two or more datasets to pinpoint individuals in datasets and reveal sensitive information. One well-known and illustrative example of these “linking attacks”, although not in a genomic context, is the linking attack that matched the entries in Netflix Prize Database and the Internet Movie Database (IMDB)20. For research purposes, Netflix released an anonymized dataset of movie ratings of thousands of viewers, which they thought was secure as the viewers’ names were removed. However, Narayanan et al20 used IMDB database, a seemingly unrelated and very large database of movie viewers, linked the two databases, and revealed identities and personal information (movie history and choices) of many viewers in the Netflix database. The fact that Netflix and IMDB host millions of individuals in their databases renders the question of detection of an individual in these database irrelevant since any random individual is very likely to be in one or both of these databases but the focus of attacks turns to matching individuals in the databases. Consequently, as the databases grow, the attacks for detection of an individual in a database become unimportant and the linking attacks become more admissible in order to characterize individuals’ sensitive information. In the genomic privacy context, as the size and number of the genotype and phenotype datasets increase, possibility of potentially linkable datasets will increase, which may make scenarios similar to Netflix attacks a reality in genomic privacy.

Several studies on genomic privacy address the linking of different datasets for re-identifying individuals and characterizing their sensitive information. Gymrek et al21 revealed the identities of several male participants of 1000 Genomes Project12 by using the short tandem repeats on Y-chromosome as an individual identifying biomarker and linking the genotypes to online genetic genealogy databases. A detailed review can be found elsewhere22. In addition, different formalisms for protecting sensitive information have been proposed and applied to genomic privacy. These censor or hide information, or aim at ensuring statistical indistinguishability of individuals in the released data. For example, differential privacy23 involves building data release mechanisms that have guaranteed bounds on the leakage of sensitive information. The release mechanisms track how much information is leaked and stops release when the estimated leakage is above a predetermined threshold. Although this approach is theoretically very appealing, it can substantially decrease the utility of the biological data24. In addition, the release mechanism must keep track of all the queries, which can cause complications in data sharing25. Homomorphic encryption26 enables performing analysis on encrypted data directly. Complete protection of sensitive information is guaranteed as the data processors never interact with the unencrypted sensitive information. The drawback, however, is high computational and storage requirements. Another well-established formalism is k-anonymization27,28. Before releasing the dataset, it is anonymized by data perturbation techniques for ensuring that no combination of features in the dataset are shared by fewer than k individuals. In this approach the anonymization process has, however, excessive computational complexity and is not practical for high dimensional biomedical datasets29. Several variants have been proposed for extending k-anonymity framework30,31. A majority of these studies aim at protecting the genomic variants and identities of individuals in databases. Different aspects of genomic privacy, pertaining linkability of high dimensional phenotype datasets to genotypes, are yet to be explored.

In our studies, we have focused on characterizability of the individuals’ sensitive information in the context of linking attacks, where the adversary exploits the genotype-phenotype correlations to link different datasets and potentially reveal sensitive information. In general, the high dimensional phenotype datasets generated in genomic studies harbor a number of phenotypes that contain sensitive information, like disease status, and other phenotypes, while not sensitive, may have subtle correlations with genomic variant genotypes. Many quantitative phenotypes can be linked to genotypes using public quantitative trait loci (QTL) datasets. Some of the high dimensional genomic quantitative traits and corresponding QTLs are gene expression levels (eQTLs), protein levels (pQTLs32,33), DNase hypersensitivity site signals (dsQTLs34), ribosome occupancy (rQTLs35), DNA methylation levels (meQTLs36), histone modification levels (haQTLs37–39), RNA splicing (sQTLs40), and also higher order traits like network modularity (modQTLs41). Other QTLs associated with single dimensional non-genomic phenotypes include body mass index42, basal glucose levels43, and serum cholesterol levels32,44. Each QTL can potentially cause a small amount of genotypic information leakage. As these QTLs are often identified and reported at genomic scale, when an adversary utilizes a large number of QTLs in the attack, he/she can accurately link the sensitive phenotypes to the genotype dataset. Since genotypes can almost perfectly identify an individual, this linking attack can potentially cause a breach of privacy for the individuals who participated in the studies.

Among all the datasets, the most abundant and well-studied genotype-phenotype correlation dataset is expression quantitative trait loci (eQTL) datasets. These datasets are generated by genome-wide screening for correlations between the variant genotypes and gene expression levels usually through RNA sequencing or expression arrays40,45,46. The eQTL datasets are especially useful in the context of linking attacks since there is a large and growing compendium of public eQTL datasets47. [For example, the GTex Project hosts a sizable set of eQTL dataset from multiple studies where the users can view in detail how the genotypes and expression levels are associated10,41. In order to demonstrate our results and build the formulations in a specific context, we have focused on eQTL datasets and linking of gene expression and genotype datasets. We are currently working on extending our analyses to other genotype-phenotype datasets.

Specifically, we have analyzed the genotype predictability and studied the tradeoff between the amount of information leakage and correct predictability of the genotypes. This enables one to quantify information leakage jointly with the predictability of genotypes from phenotype data. We also developed a basic 3 step individual characterization framework, which can enables systematic analysis of any linking attack. We studied different aspects of vulnerability using the framework. We also developed practical linking attacks. In one scenario, we showed how outliers of gene expression levels can be utilized for pinpointing individuals. We are evaluating the extent and practicality of these attacks on different genotype and phenotype datasets.

# IMPLICATIONS AND FUTURE DIRECTIONS OF OUR WORK

Increasing pace of data generation and the policies to encourage genomic data sharing will make genomic privacy a topic of hot debate. In the analysis of genomic privacy, it is necessary to consider the basic premise of sharing any type of personal information: There is always an amount of leakage in the sensitive information48. In addition, as shown by previous studies, we often cannot propose black-and-white solutions to problems in privacy which mainly roots from the multifaceted nature of privacy. We believe these make it necessary for the genomic data sharing and publishing mechanisms to incorporate statistical quantification methods before the datasets are released. This is recently recently recognised49. Legislative decision making processes should incorporate the quantified risk estimates of leakage as an objective factor. The quantification methodology and the analysis frameworks presented in this study can be applied for analysis of the information leakage in the datasets where the correlative relations between datasets can be exploited for performing linking attacks. In accordance to a utility policy, the leakage risk can be evaluated against the utility requirements so as to assess the suitability of different data release mechanisms.

# REFERENCES

1. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12,** 125 (2011).

2. Rodriguez, L. L., Brooks, L. D., Greenberg, J. H. & Green, E. D. The Complexities of Genomic Identifi ability. *Science (80-. ).* **339,** 275–276 (2013).

3. Sweeney, L., Abu, A. & Winn, J. Identifying Participants in the Personal Genome Project by Name. *SSRN Electron. J.* 1–4 (2013). doi:10.2139/ssrn.2257732

4. infographic-printable.pdf. at <http://www.nih.gov/precisionmedicine/infographic-printable.pdf>

5. Collins, F. S. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372,** 793–795 (2015).

6. Plan for Increasing Access to Scientific Publications - NIH-Public-Access-Plan.pdf. at <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>

7. GENOMIC DATA SHARING (GDS) Home. at <http://gds.nih.gov/index.html>

8. Sweeney, L. *Uniqueness of Simple Demographics in the U.S. Population, LIDAP-WP4*. *Forthcom. B. entitled, Identifiability Data.* (2000).

9. Golle, P. Revisiting the uniqueness of simple demographics in the US population. in *Proc. 5th ACM Work. Priv. Electron. Soc.* 77–80 (2006). doi:http://doi.acm.org/10.1145/1179601.1179615

10. Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

11. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

12. The 1000 Genomes Project Consortium. An integrated map of genetic variation. *Nature* **135,** 0–9 (2012).

13. Collins, F. S. The Cancer Genome Atlas ( TCGA ). *Online* 1–17 (2007).

14. Pakstis, A. J. *et al.* SNPs for a universal individual identification panel. *Hum. Genet.* **127,** 315–324 (2010).

15. Wei, Y. L., Li, C. X., Jia, J., Hu, L. & Liu, Y. Forensic Identification Using a Multiplex Assay of 47 SNPs. *J. Forensic Sci.* **57,** 1448–1456 (2012).

16. Church, G. *et al.* Public access to genome-wide data: Five views on balancing research with privacy and protection. *PLoS Genet.* **5,** (2009).

17. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nat. Rev. Genet.* **9,** 406–411 (2008).

18. Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** (2008).

19. 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).

20. Narayanan, A. & Shmatikov, V. Robust de-anonymization of large sparse datasets. in *Proc. - IEEE Symp. Secur. Priv.* 111–125 (2008). doi:10.1109/SP.2008.33

21. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339,** 321–4 (2013).

22. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15,** 409–21 (2014).

23. Dwork, C. Differential privacy. *Int. Colloq. Autom. Lang. Program.* **4052,** 1–12 (2006).

24. Fredrikson, M. *et al.* Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. in *23rd USENIX Secur. Symp.* 17–32 (2014). at <http://www.biostat.wisc.edu/~page/WarfarinUsenix2014.pdf>

25. Adam, N. R. & Worthmann, J. C. Security-control methods for statistical databases: a comparative study. *ACM Comput. Surv.* **21,** 515–556 (1989).

26. Gentry, C. A FULLY HOMOMORPHIC ENCRYPTION SCHEME. *PhD Thesis* 1–209 (2009). doi:10.1145/1536414.1536440

27. SWEENEY, L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.* **10,** 557–570 (2002).

28. Loukides, G., Gkoulalas-Divanis, A. & Malin, B. Anonymization of electronic medical records for validating genome-wide association studies. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 7898–7903 (2010).

29. Meyerson, A. & Williams, R. On the complexity of optimal K-anonymity. in *Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst.* 223–228 (2004). doi:10.1145/1055558.1055591

30. Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. L-diversity. *ACM Trans. Knowl. Discov. Data* **1,** 3–es (2007).

31. Ninghui, L., Tiancheng, L. & Venkatasubramanian, S. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. in *Proc. - Int. Conf. Data Eng.* 106–115 (2007). doi:10.1109/ICDE.2007.367856

32. Holdt, L. M. *et al.* Quantitative trait loci mapping of the mouse plasma proteome (pQTL). *Genetics* **193,** 601–608 (2013).

33. Stark, A. L. *et al.* Protein Quantitative Trait Loci Identify Novel Candidates Modulating Cellular Response to Chemotherapy. *PLoS Genet.* **10,** (2014).

34. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).

35. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. *Science (80-. ).* **347,** 664–667 (2014).

36. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12,** R10 (2011).

37. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Sci. (New York, NY)* **342,** 747–749 (2013).

38. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342,** 744–7 (2013).

39. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Sci. (New York, NY)* **342,** 750–752 (2013).

40. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

41. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-. ).* **348,** 648–660 (2015).

42. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42,** 937–948 (2010).

43. Cheverud, J. M. *et al.* Quantitative trait loci for obesity- and diabetes-related traits and their dietary responses to high-fat feeding in LGXSM recombinant inbred mouse strains. *Diabetes* **53,** 3328–3336 (2004).

44. Beekman, M. *et al.* Evidence for a QTL on chromosome 19 influencing LDL cholesterol levels in the general population. *Eur. J. Hum. Genet.* **11,** 845–850 (2003).

45. Stranger, B. E. *et al.* Patterns of Cis regulatory variation in diverse human populations. *PLoS Genet.* **8,** (2012).

46. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).

47. Xia, K. *et al.* SeeQTL: A searchable database for human eQTLs. *Bioinformatics* **28,** 451–452 (2012).

48. Narayanan, A. *et al.* *Redefining Genomic Privacy: Trust and Empowerment*. *bioRxiv* (2014). doi:10.1101/006601

49. US Department of Commerce, N. NIST Requests Comments on a Draft Privacy Risk Management Framework. at <http://www.nist.gov/itl/201506\_privacy\_framework.cfm>