Identifying Anonymous Donors of Genetic Information

Mohammed Alser[1], Nour Almadhoun[2], Azita Nouri[3], Can Alkan[4], and Erman Ayday[5]
Computer Engineering Department, Bilkent University, 06800, Ankara, Turkey
{[1]mohammedalser, [2]nour.madhoun, [3]azita}@bilkent.edu.tr, {[4]calkan, [5]erman}@cs.bilkent.edu.tr

The rapid progress in today's genome sequencing technologies leads to availability of high amounts of genomic data for as little as few hundred dollars. This provides an adequate basis for several important applications and studies. Genomic research typically includes collecting samples from thousands of individuals [1]. Furthermore, a large push is underway to sequence hundreds of thousands to millions of genomes aiming to discover the functional impact of *de novo* (not inherited from either parent) genetic variations on diseases such as autism and cancer [2, 3]. Accelerating the pace of biomedical breakthroughs and discoveries also necessitates granting open access to the genetic databases. This trend has caused the launch of more than one thousand freely available online genetic databases worldwide, in which individuals publicly share their genomic data [4]. The public in different countries (USA, Sweden, Japan, and Singapore) have positive attitude towards their willingness to donate genetic samples to support the personalized medicine studies [5-11]. However, one growing concern is the ability to protect the privacy of the sensitive information and its owner. Thus, the biggest challenge of widely utilizing the human genomes and pushing the frontiers of the genetics research is social, and not technical [12]. In this work, we survey and categorize a wide spectrum of known privacy breaching strategies to human genomic data as follows.

Meta-data & side-channel leaks: The curious party needs both human genomic data, which is already anonymized and available online, and additional metadata, such as basic demographic details, pedigree structure, voter list, or medical reports [13, 14]. Once the owner of a genome is identified, he is faced with the risk of genetic discrimination, financial loss, and blackmail.

Genealogical triangulation: The adversary can take advantage of the correlation between the Y-chromosome and surname, and compare the Y-haplotype of the unknown genome to haplotype records in genealogy databases [15]. The power of this attack stems from exploiting information from distant patrilineal relatives of the unknown's genome. Surnames are also highly searchable through public records and social networks. In 2013, a study [16] showed that five successful surname inferences lead to exposition of the identity of nearly 50 members of three families, who might have no acquaintance with the person who released his genetic data.

Phenotypic prediction: Visible phenotypes with high heritability can be derived from genetic data (e.g., eye color, facial morphology, age prediction) could serve as quasi-identifiers [17, 18]. This technique depends on reducing the degree of uncertainty to predict the identity with the help of public records and social networks.

Disclosure attacks via DNA: When the adversary gains access to the DNA sample of the target, by using the identified DNA, he or she can search genetic databases with sensitive attributes (e.g., drug abuse). Matching the identified DNA with the database reveals the link between the person and the sensitive attribute [19-22].

Completion attacks: Genotype imputation is a well-studied task where the genetic information of a known individual can be reconstructed/predicted from partial data by completing the missing genotype values [23, 24].

After this categorization, in the remaining of this work, we also discuss potential countermeasure mechanisms for the above threats.

References

[1] F. S. Collins, *et al.*, "A vision for the future of genomics research," *Nature*, vol. 422, pp. 835-847, 2003.

[2] I. Iossifov, *et al.*, "The contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, pp. 216-221, 2014.

[3] G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56-65, 2012.

[4] M. Y. Galperin, *et al.*, "The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection," *Nucleic acids research*, vol. 43, pp. D1-D5, 2015.

[5] E. Kobayashi and N. Satoh, "Public involvement in pharmacogenomics research: a national survey on public attitudes towards pharmacogenomics research and the willingness to donate DNA samples to a DNA bank in Japan," *Cell and tissue banking*, vol. 10, pp. 281-291, 2009.

[6] J. M. Pulley, *et al.*, "Attitudes and perceptions of patients towards methods of establishing a DNA biobank," *Cell and tissue banking*, vol. 9, pp. 55-65, 2008.

[7] C. L. Storr, *et al.*, "Genetic research participation in a young adult community sample," *Journal of community genetics*, vol. 5, pp. 363-375, 2014.

[8] Å. Kettis-Lindblad, *et al.*, "Genetic research and donation of tissue samples to biobanks. What do potential sample donors in the Swedish general public think?," *The European Journal of Public Health*, vol. 16, pp. 433-440, 2006.

[9] I. Ishiyama, *et al.*, "Relationship between public attitudes toward genomic studies related to medicine and their level of genomic literacy in Japan," *American Journal of Medical Genetics Part A*, vol. 146, pp. 1696-1706, 2008.

[10] K. Hoeyer, *et al.*, "Informed consent and biobanks: a population-based study of attitudes towards tissue donation for genetic research," *Scandinavian Journal of Public Health*, vol. 32, pp. 224-229, 2004.

[11] A. K. Rahm, *et al.*, "Biobanking for research: a survey of patient population attitudes and understanding," *Journal of community genetics*, vol. 4, pp. 445-450, 2013.

[12] G. Neff, "Why big data won't cure us," *Big data*, vol. 1, pp. 117-123, 2013.

[13] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, pp. 557-570, 2002.

[14] L. Sweeney, *et al.*, "Identifying participants in the personal genome project by name," *Available at SSRN 2257732*, 2013.

[15] T. E. King and M. A. Jobling, "What's in a name? Y chromosomes, surnames and the genetic genealogy revolution," *Trends in Genetics*, vol. 25, pp. 351-360, 2009.

[16] M. Gymrek, *et al.*, "Identifying personal genomes by surname inference," *Science*, vol. 339, pp. 321-324, 2013.

[17] P. Directive, "Identifiability in genomic research," 2007.

[18] M. Kayser and P. de Knijff, "Improving human forensics through advances in genetics, genomics and molecular biology," *Nature Reviews Genetics*, vol. 12, pp. 179-192, 2011.
[19] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, pp. 409-421, 2014.
[20] H. K. Im, *et al.*, "On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy," *The American Journal of Human Genetics*, vol. 90, pp. 591-598, 2012.
[21] T. Lumley and K. Rice, "Potential for revealing individual-level information in genome-wide association studies," *JAMA*, vol. 303, pp. 659-660, 2010.
[22] A. J. Pakstis, *et al.*, "SNPs for a universal individual identification panel," *Human genetics*, vol. 127, pp. 315-324, 2010.
[23] D. R. Nyholt, *et al.*, "On Jim Watson's APOE status: genetic information is hard to hide," *European Journal of Human Genetics*, vol. 17, p. 147, 2009.
[24] J. Kaiser, "Agency nixes deCODE's new data-mining plan," *Science*, vol. 340, pp. 1388-1389, 2013.