COMMENTARY

On genomics, kin, and privacy [version 1; referees: 3 approved]

Amalio Telenti[1], Erman Ayday[2], Jean Pierre Hubaux[2]

[1]Department of Laboratories, University Hospital of Lausanne, Lausanne, Switzerland
[2]Laboratory for Communications and Applications Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

v1 **First published:** 31 Mar 2014, **3**:80 (doi: 10.12688/f1000research.3817.1)
Latest published: 31 Mar 2014, **3**:80 (doi: 10.12688/f1000research.3817.1)

Abstract

The storage of greater numbers of exomes or genomes raises the question of loss of privacy for the individual and for families if genomic data are not properly protected. Access to genome data may result from a personal decision to disclose, or from gaps in protection. In either case, revealing genome data has consequences beyond the individual, as it compromises the privacy of family members. Increasing availability of genome data linked or linkable to metadata through online social networks and services adds one additional layer of complexity to the protection of genome privacy. The field of computer science and information technology offers solutions to secure genomic data so that individuals, medical personnel or researchers can access only the subset of genomic information required for healthcare or dedicated studies.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 31 Mar 2014	✓ report	✓ report	✓ report

1 **XiaoFeng Wang**, University of Indiana at Bloomington USA

2 **Xiaoqian Jiang**, University of California, San Diego USA

3 **Florian Kerschbaum**, SAP Germany

Discuss this article

Comments (0)

Corresponding authors: Amalio Telenti (amalio.telenti@chuv.ch), Jean Pierre Hubaux (jean-pierre.hubaux@epfl.ch)

How to cite this article: Telenti A, Ayday E and Hubaux JP. **On genomics, kin, and privacy [version 1; referees: 3 approved]** *F1000Research* 2014, **3**:80 (doi: 10.12688/f1000research.3817.1)

Grant information: A.T. is funded by the Swiss National Science Foundation (SNF #141234 and CRSII3_147665).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors hold a patent on protecting the privacy of genomic data in medical tests using cryptographic techniques.

First published: 31 Mar 2014, **3**:80 (doi: 10.12688/f1000research.3817.1)

Introduction

The recent authorization of a sequencing platform for clinical use by the Food and Drug Administration will expand and accelerate the use of genetic information in medical care[1]. Progress is particularly impressive in the deployment of sequencing tools for neonatal diagnostics[2]. Commoditization of genome-wide genotyping and sequencing is happening as rapidly outside of the medical setting – prominently through companies offering "direct to consumer" (DTC) services. There is full awareness of the need to protect these data[1] – while simultaneously supporting their use in research[3]. Here, we discuss how protection of genome data from medical and non-medical sources needs to be reframed considering the mutual implications of personal decision, online social networks and consequences to relatives.

On personal decisions

Paradoxically, genomics is an attractive field for individual or collective altruism – many people are willing to place their genome data in the public domain, and to actively engage in genomic research. The academic community is also calling for definitive actions to support global data-sharing[3]. Many research participants count on the protection of their identity. However, current strategies have proven insufficient to stop sophisticated attacks on genetic data. A recent study[4] demonstrated the feasibility of re-identifying DNA donors from a public research database by using information available from popular genealogy websites. Attackers can also take advantage of gaps in the protection of other sources of data, for example census and voter lists, hospital insurance reports, and increasingly, from online social networks (see below). Genome data in the wrong hands could have undesirable consequences: from discrimination, or release of paternity, ancestry or other data that the participant did not intend to be public, to more prosaic usages such as targeted advertisements based on genome information.

Genome and online social networks

Online social platforms are convenient sites for posting data but they are susceptible to "multilayer attacks": the possibility to simultaneously aggregate data from online social networks (e.g., Facebook), health related websites (e.g., patientslikeme.com), platforms for sharing genome data (e.g., OpenSNP.org), family history resources (e.g., ancestry.com), research datasets (e.g., 1000 Genomes Project), and public records (e.g., voter registration forms) can help an attacker de-anonymize the owner of an anonymized genome and/or infer the genomic data of his/her family members. We illustrate in Figure 1A the feasibility and ease of cross-identification of a given individual across various genetic and non-genetic platforms, including the reconstitution of parts of the family pedigree.

On kinship issues

Kin aspects of genomics were well publicized by the recent controversy regarding the public release of the genome of Henrietta Lacks (August 1, 1920 – October 4, 1951). HeLa, a cell line established from Lacks, has been used for decades in research laboratories world-wide. Recently, HeLa cells were sequenced and the genome data posted online without the consent of her relatives, who subsequently complained that this accounted to revealing private information about the family. The multilayer attacks mentioned above can reconstruct phylogenies from revealed genomes and open the door to genetic prediction of family members. The amount of kin privacy lost from such attacks can be precisely estimated (Figure 1B). As more individuals will have their genome sequenced or genotyped in coming years, the loss of privacy of family members through multilayer attacks will increase if no action is taken.

Solutions from computer science

There is little doubt that genome privacy will be challenged – in particular if the medical establishment relies solely on legal deterrents and conventional protection of stored data, or if it resorts to ineffective deidentification and anonymization of genome data shared for the purpose of research. However, personal genetic tests and genomic research are possible without jeopardizing the genomic privacy of the individual or of family members. In particular, IT security provides a trove of solutions. These include using efficient cryptographic techniques for privacy-preserving personalized medicine[5,6], and for genomic research[7]. With such approaches, genomic data are always stored in encrypted form and medical personnel or researchers can access only the subset of genomic information required for healthcare or dedicated studies. Similarly there are obfuscation-based solutions[8] to use genomic data in research settings in a privacy-preserving way.

Some genome researchers may be tempted to belittle the threat raised by the possible leakage of genomic data. This is a mistake, because progress in genetics is likely to make these data more and more meaningful. In addition, if it appears that genomic data are not properly protected, people could start distrusting genetics, with negative consequences for the progress of medicine. Protection needs to consider both the interest of the individual and of relatives. It is important to learn from errors in Internet security over the last decades. In that field, tools and solutions are often lagging behind threats.

The first meeting exclusively dedicated to genomic privacy took place in October 2013 at the Leibniz Center for Informatics in Dagstuhl, Germany (http://www.dagstuhl.de/13412). As one of the outcomes, the community set up a web site reporting the efforts and progress on this topic: https://genomeprivacy.org/. Notably, this site contains the list of research groups active in this field, as well as basic information to facilitate the understanding of this novel field. It is our conviction that by pooling together the skills of geneticists, law scholars, ethicists and computer scientists, we are still in time to strike an appropriate balance between accessibility to genome data and their protection.

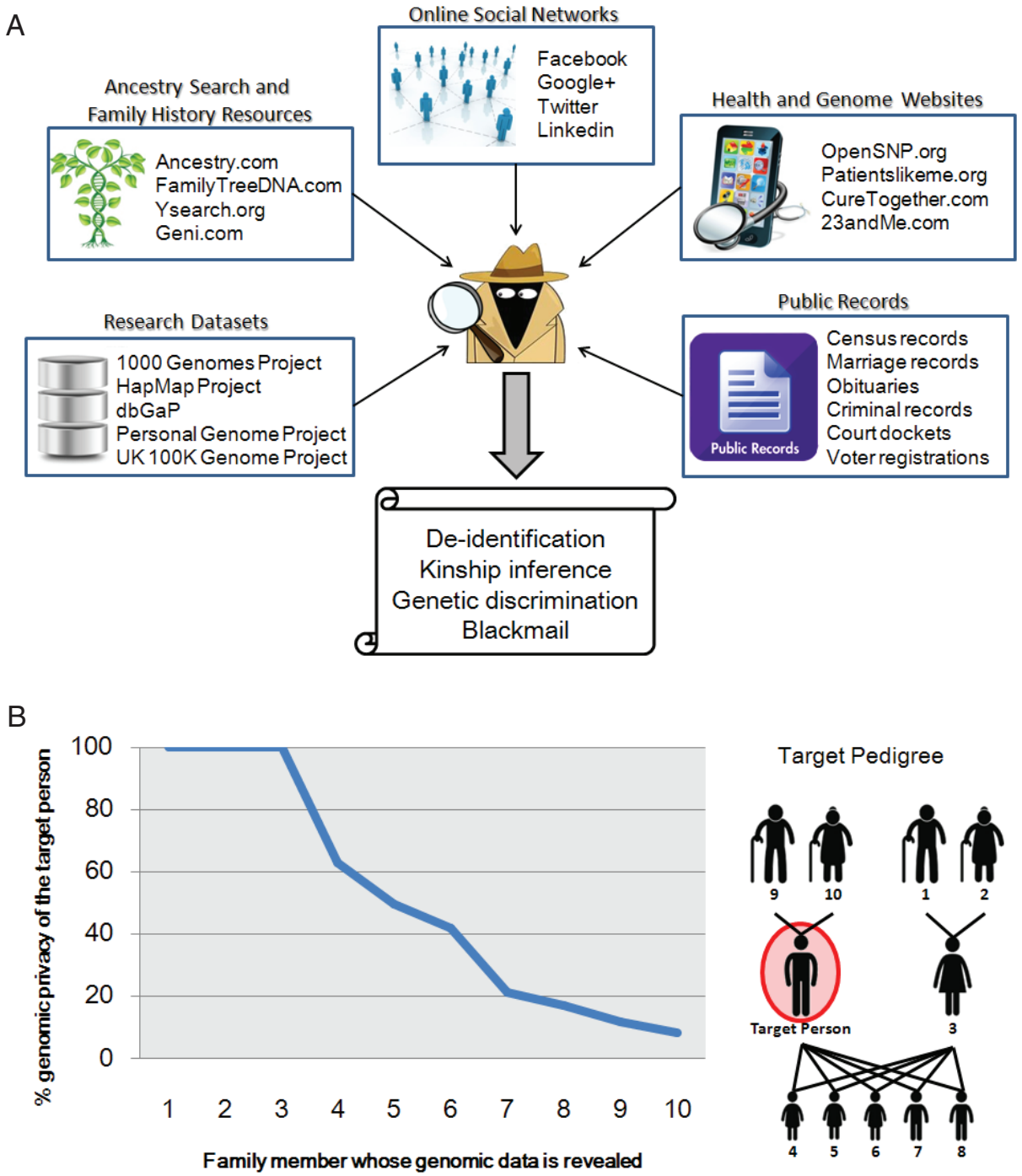

Figure 1. Attacks on genomic privacy. (**A**) Multilayer attacks using data from genomic and non-genomic platforms. An attacker can obtain the anonymized genomic data of an individual from one of the genome data websites (e.g., openSNP.org). Then, the attacker can de-anonymize the owner of the genome (i.e., learn his/her identity) by matching his/her phenotypic, demographic and administrative information (e.g., profile picture, age, gender, ZIP code) across the individual's online social network profile. Once the individual is de-identified, the attacker can also determine his/her family members from a family history resource (e.g., ancestry.com) and infer the genomic data of family members from the individual's retrieved genome. For example, owners of some genomes uploaded to openSNP can be de-anonymised using their Facebook profiles. For 6 individuals who publicly revealed the names of some of their relatives on Facebook, 29 familial relationships could be identified[9]. (**B**) Decrease in genomic privacy of the target person (circled in red) when the genomes of his family members are gradually revealed. The health privacy of family members can be quantified. For example, two single nucleotide polymorphisms (rs7412 and rs429358) of the Apolipoprotein E (*ApoE*) gene are associated with increased risk for Alzheimer's disease. The identification in several members of the pedigree of a carrier status for those risk alleles can reveal the *ApoE4* status of the target person to the attacker.

Author contributions

A.T., E.A. and H.-P. H. conceived the content of the commentary and wrote the paper.

Competing interests

The authors hold a patent on protecting the privacy of genomic data in medical tests using cryptographic techniques.

Grant information

A.T. is funded by the Swiss National Science Foundation (SNF #141234 and CRSII3_147665).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Collins FS, Hamburg MA: **First FDA authorization for next-generation sequencer.** *N Engl J Med.* 2013; **369**(25): 2369–2371.
PubMed Abstract | Publisher Full Text
2. Yang Y, Muzny DM, Reid JG, *et al.*: **Clinical whole-exome sequencing for the diagnosis of mendelian disorders.** *N Engl J Med.* 2013; **369**(16): 1502–1511.
PubMed Abstract | Publisher Full Text
3. Hayden EC: **Geneticists push for global data-sharing.** *Nature.* 2013; **498**(7452): 16–17.
PubMed Abstract | Publisher Full Text
4. Gymrek M, McGuire AL, Golan D, *et al.*: **Identifying personal genomes by surname inference.** *Science.* 2013; **339**(6117): 321–324.
PubMed Abstract | Publisher Full Text
5. Ayday E, Raisaro JL, Rougemont J, *et al.*: **Protecting and evaluating genomic privacy in medical tests and personalized medicine**. ACM Workshop on Privacy in the Electronic Society (WPES 2013), Berlin, Germany, 2013.
Publisher Full Text
6. Baldi P, Baronio R, De Cristofaro E, *et al.*: **Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes**. ACM Conference on Computer and Communications Security (CCS), 2011.
Publisher Full Text
7. Kantarcioglu M, Jiang W, Liu Y, *et al.*: **A cryptographic approach to securely share and query genomic sequences.** *IEEE Trans Inf Technol Biomed.* 2008; **12**(5): 606–617.
PubMed Abstract | Publisher Full Text
8. Johnson A, Shmatikov V: **Privacy-preserving data exploration in genome-wide association studies**. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013.
Publisher Full Text
9. Humbert M, Ayday E, Hubaux JP, *et al.*: **Addressing the concerns of the lacks family: quantification of kin genomic privacy**. ACM Conference on Computer and Communications Security (CCS), 2013.
Publisher Full Text

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 29 August 2014

doi:10.5256/f1000research.4089.r5895

Florian Kerschbaum
SAP, Karlsruhe, Germany

This article raises a very important issue: the difficulty of providing privacy for genetic information in the light of inheritance. I cannot stress enough how important this aspect is, since it requires data protection measures, such as the mentioned cryptographic or obfuscation-based approaches, to be utterly restrictive. The article gives clear examples where -- intentionally or unintentionally -- leaked information allowed harmful inferences. The authors combine information from public genomes and social networks to infer information about people who have not released any information about their genes. These examples should be taken very, very seriously, since they are raised only at the very beginning of the scientific development. As the paper argues we can expect the use of genomics to significantly grow. Genetic information can prove at least as harmful as location information provided by cell phones, but it is impossible (for now) to change it. It is therefore a scientific and societal challenge to protect genomic information much better than we protect our information in current telecommunication networks. The article references excellent works from the computer security community that can certainly provide a guiding direction, but even these mechanisms need to necessarily leak some information about the genomes. I hope that medical and computer science researchers will take the challenge described by the article seriously and look for mechanisms that control the entire information in all medical or non-medical information system based on direct or indirect genomic data.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 29 August 2014

doi:10.5256/f1000research.4089.r5889

Xiaoqian Jiang
Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA

This is a timely commentary on privacy, kin, and genomics. Today, many gene donators are still ignorant of the potential impact of information leakage to the family when their genome data are made public. This

problem is becoming more critical as the younger generation reveals themselves and family members on online social networks and open family history resources made it possible to link individuals. I found this topic to be extremely important.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 14 April 2014

doi:10.5256/f1000research.4089.r4304

XiaoFeng Wang
Centre for Security Informatics, School of Informatics and Computing, University of Indiana at Bloomington, Bloomington, IN, USA

This paper discusses the challenge of protecting human genome data, particularly its unique feature in that one's DNA data can be used to infer the private health information of those genetically related to them. The authors talk about the conflict between the perception that the decision on releasing one's DNA materials is personal, and how it can actually impact on the privacy of their kin. They further sketch a technique for quantifying such an information leak, and demonstrate that the threat is realistic, given the de-anonymization attack that can happen through the booming online social networks. I feel that this article provides useful information for raising the awareness of the uniqueness and significance of genome privacy. This, hopefully, will lead to a broad, in-depth conversation among genomics researchers, security and privacy researchers, bioethics experts, genomics industry, policy makers and the public on how to effectively regulate the dissemination of human DNA data to facilitate scientific research, without undermining DNA donors' privacy and well-being.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.