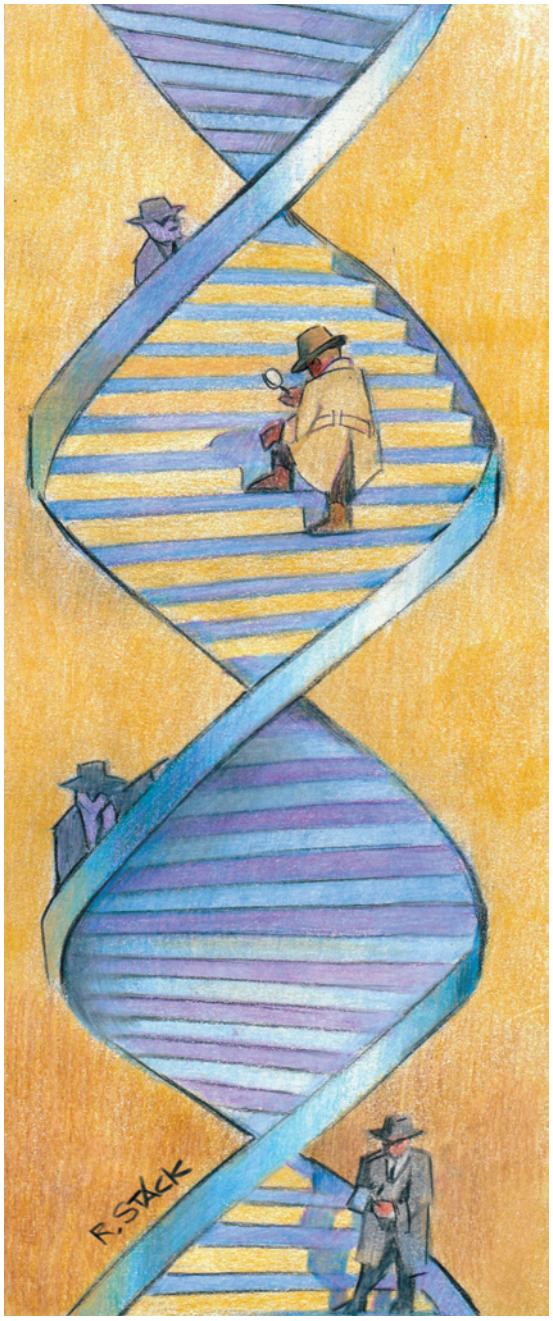
Editors: Angela Sasse, A.Sasse@cs.ucl.ac.uk | Alessandro Acquisti, acquisti@andrew.cmu.edu

Genomic Privacy and the Rise of a New Research Community

Emiliano De Cristofaro | University College London



Over the past two decades, advances in whole genome sequencing (WGS)—the suite of technologies used to determine an organism's complete DNA sequence—have been quite exceptional, boasting a drop in costs and an improvement in throughput significantly faster than what Moore's law would predict. The first full sequencing of a human genome was completed in 2003 after a 13-year, US$3 billion-worth research collaboration (the Human Genome Project), which involved more than 20 institutions worldwide. Since then, numerous companies and research institutions have competed in a race toward more affordable and accurate technologies, with prices plummeting to hundreds of thousands of dollars by 2008. The long-anticipated $1,000 threshold was breached by the San Diego–based company Illumina early this year.

Inexpensive WGS facilitates the collection of a large number of digitized genomes, thus, it's considered a key enabler of research in genomics. In fact, it's essential for genome-wide association studies (GWAS), which rely on very large datasets to discover correlations between common genetic variants and traits, for example, disease predisposition or response to treatment.

As the cost of full sequencing drops below that of an MRI scan, genome sequencing will soon be available to the masses, enabling the advent of a new era of personalized medicine in which medical care can be tailored to an individual patient's genetic makeup. The personalized medicine paradigm relies on the ability to determine individuals' unique genetic characteristics to more effectively diagnose diseases, select treatments that increase the chances of a successful outcome, and reduce possible adverse reactions. Genomic tests are already routinely used to predict patients' response to several drugs and to help doctors assess the right therapy for millions of HIV, cancer, and leukemia patients. Naturally, the availability of a patient's fully sequenced genome will only make it easier for clinicians to run complex tests in a matter of seconds using specialized computational algorithms (as opposed to more expensive and slower in vitro tests), thus markedly facilitating a personalized and participatory approach to healthcare. At the same time, however, progress in WGS raises significant challenges to privacy due to the unprecedented sensitivity of genomic data.

The Greater Good and the Need for Privacy

Rapid developments in genomics are often promoted as being dependent on resources that facilitate and encourage sharing of sequenced genomes, making them accessible to researchers for different research purposes. Funding agencies have actually introduced requirements

Copublished by the IEEE Computer and Reliability Societies

that data sharing must be considered in all genomics-related grant applications. Clearly, sharing is an important asset for hypothesis-driven research. In fact, to robustly derive conclusions on disease susceptibility from genetic features, experiments may require the availability of millions of DNA samples. Even if sequencing costs drop below $1,000, it would still be unrealistic to assume that each research team could afford to recruit and sequence the genome of that many participants.

Recent public initiatives, such as the Personal Genome Project (www.personalgenomes.org), aim to create public datasets for research purposes, involving volunteers who agree to have their genome sequenced and made publicly available, along with other personal information, for "the greater good." The UK recently announced an investment of £100 million and a plan to fully sequence the genome of 100,000 patients by 2017.

Unsurprisingly, genomic data sharing and dissemination prompt important privacy, ethical, and legal concerns. The human genome not only uniquely and irrevocably identifies its owner, but it also contains information about ethnic heritage, predisposition to diseases (such as Alzheimer's, breast and ovarian cancer, and so on) and conditions (including mental disorders, such as schizophrenia), and many other phenotypic traits.

Because the human genome contains detailed information about ethnicity and susceptibility to somatic and mental conditions, its disclosure is often associated to the fear of eugenism—genetic discrimination—which bears potentially dreadful implications for social dynamics as well as hiring and health insurance practices.

As noted in a recent article by Erman Ayday and colleagues, although some recourse is possible in more traditional privacy leakage scenarios (for example, stolen passwords can be changed), the consequences of genomic data disclosure aren't limited in time, and genomes can't be "revoked."[1] Ayday and colleagues also argue that privacy threats posed by access to large numbers of digitized genomes aren't comparable with those arising from an attacker collecting a sample such as shed hair or saliva traces from a glass and surreptitiously sequencing the victim's genome. Such an attack poses a credible threat only against a target individual or a small group of people, and also incurs a remarkably higher cost.

Due to its hereditary nature, disclosing one's genome essentially implies disclosing the genomes of close relatives, as demonstrated, among others, in a recent paper by Mathias Humbert and colleagues.[2] The family of Henrietta Lacks—a cancer patient whose genome was sequenced and published online without her or her family's consent—reached a legal settlement with US federal authorities on the basis of the violation of their privacy due to the disclosure of Ms. Lack's genome. Masking sensitive portions of the genome—such as mutations that indicate predisposition to Alzheimer's disease—is almost impossible because the correlation (specifically, linkage disequilibrium) between one or multiple genetic mutations can often be used to reconstruct the "redacted" features.

Concerns about privacy and genetic discrimination are emerging from ethnographic studies involving prospective participants of sequencing programs as well. For instance, experimental studies by Amy McGuire and colleagues revealed a distinct correlation between privacy fears and opting out of sequencing programs.[3] Consequently, successfully addressing privacy concerns becomes a crucial step needed to enable genomic research relying on donors' genomes.

Current legislation—such as the 2003 Health Insurance Portability and Accountability Act (HIPAA) or the Genetic Information Nondiscrimination Act (GINA)—establishes a frame of guidelines and policies, but it doesn't technically enforce effective requirements on safe and secure storage/processing of sequenced genomes. Protection of human subject research is triggered by the identifiability of data, thus, researchers typically attempt to ensure participants' anonymity via data de-identification: stripping name, date of birth, photographs, or other identifiers. However, in the context of whole genomes, de-identification is hardly effective, because the genome is by nature a unique identifier, as demonstrated by Melissa Gymrek and colleagues, who were able to reidentify DNA donors from a public research database using information available from popular genealogy websites.[4] A recent article by Yaniv Erlich and Arvind Narayanan provides a systematic overview of the different routes for breaking genetic privacy.[5] Breaches can be triggered by

- identity tracing, that is, exploiting quasi-identifiers in DNA data to uncover an unknown genetic dataset's identity;
- attribute disclosure, that is, using known DNA data to link a person's identity with a sensitive phenotype; and
- completion techniques, that is, uncovering sensitive genomic areas that were masked.

Nonetheless, the report released in late 2012 by the US Presidential Commission for the Study of Bioethical Issues (http://bioethics.gov/node/764) actually recommends de-identification to protect privacy, even though, as discussed

earlier, this is completely moot for whole genomes.

Regulations also require that, prior to sequencing, investigators obtain the subjects' informed consent, and that participants know who owns the data, who guarantees proper handling, who will have further access to the data, and what security measures are in place. However, as pointed out by Deborah Mascalzoni and colleagues, this proves to be challenging against a background in which the research questions, as well as the information that can be extracted or inferred from genomes, rapidly change.[6] Restrictive interpretations of informed consent could limit the use of participants' specimens and strengthen privacy guarantees but at the potential cost of hindering genomic research (for instance, genomic datasets could not be reused to study a different disease without renewed consent). Also, similarity of related individuals' genomes raises doubts as to whether relatives should also provide consent.

WGS for the Masses: Challenges and Implications

In the not-so-distant future, cheap WGS will give the public the means to have their genome fully sequenced. Companies such as Illumina already offer commercial WGS products for private individuals who, for a few thousand dollars, receive both raw and annotated data as well as access to healthcare professionals to discuss findings and results of several genetic tests.

Arguably, we're witnessing the emergence of a novel genomics market stimulated both by medical reasons and personal curiosity, where direct-to-consumer products find an increasingly large space. For instance, between 2006 and 2013, 23andme offered a $99 assessment of inherited traits, genealogy, and congenital risk factors by genotyping saliva samples posted via mail. Ancestry.com offers a similar low-cost service, providing genealogical DNA tests, again based on genotyping saliva samples, while genepartner.com provides matchmaking services based on potential partners' genetic features.

The availability of affordable WGS will likely allow individuals to obtain, own, and retain a copy of their sequenced genome, which has implications for security professionals. Due to the sensitivity of genomic information, one open challenge is where and how a digitized genome should be stored. Ideally, individuals who request sequencing should own the result, as is already the case with other personal medical results and information. However, if we let users retain and store their sequenced genome on their computer or on a dedicated portable device, it isn't clear how to guarantee seamless access to data for medical and/or research purposes or secure storage: non-tech-savvy users are notoriously bad at maintaining secrets and dealing with cryptographic keys.

Clearly, the long-term sensitivity of genomic data, along with the ineffectiveness of de-identification and data sanitization, requires novel technical solutions.

On the other hand, if we rely on outsourcing the genomes' storage and processing to dedicated providers, we run into obvious trust, security, and privacy issues. Homomorphic encryption might seem like a "Holy Grail" solution, enabling computation over encrypted data by third-party servers storing and processing encrypted data without learning the corresponding plaintexts. However, besides its excessive complexity, computation over encrypted data might not be well-suited to a genomic setting, where computational algorithms need to cope with frequent sequencing errors and insertions or deletions. Furthermore, even if genomes are encrypted, cryptosystems considered strong today are likely to be less so in the long term, while genome sensitivity doesn't dissipate over time—especially when considering that leakage of an individual's genome affects the privacy of that person's living progeny.[1]

Even if security and privacy issues aren't yet obstructing the WGS revolution, it isn't too early to start investigating and tackling them. As often happens, one possible direction is for the security community to look back at protection mechanisms used in other contexts and take advantage of the lessons learned in recent years. The advent of new security and privacy threats such as identity theft, phishing, or the deanonymization of public datasets (such as Netflix's) often generates a lag between harm to individuals and the deployment of technical/policy countermeasures, leaving behind practical, economic, and moral damages, as well as lawsuits. Security and privacy practitioners involved with genomic testing infrastructures should be even more cautious and vocal with respect to threats to genomic privacy in light of their complexity, legal implications, and potential impact on society.

This also creates the need for educating users, developers, and policymakers about the risks and benefits of genomic data sharing and disclosure. Clearly, the long-term sensitivity of genomic data,

along with the ineffectiveness of de-identification and data sanitization, requires novel technical solutions as well as effective policy efforts, but it also creates exciting opportunities for the security community to be part of this innovation.

A New, Multidisciplinary Research Community

Reconciling privacy with progress in human genomics motivates and encourages research efforts toward secure environments for genome testing. This raises technical and ethical challenges that can't be addressed by security research alone—rather, they require the expertise and collaboration of geneticists, clinicians, biologists, bioinformaticians, ethnographers, and computer scientists.

Experts from different fields have already started to join forces, paving the way for the establishment of a new and heterogeneous community. The past few months have included several workshops and meetings (for example, the October 2013 "Genomic Privacy" seminar in Dagstuhl, Germany) that bring together practitioners and researchers from various areas to exchange results, insights, practical requirements, and ethical and legal implications related to genomic data protection.

Obviously, the brief history of such a young community features success stories, glitches, and hurdles. Collaborative efforts have started to produce encouraging feasibility results as to how to realize privacy-respecting versions of computational genomic tests. For instance, the work by Pierre Baldi and colleagues (a team composed of a bioinformatics expert, a biologist, and three security researchers) highlights the need for combining domain knowledge in genomics and security and analyzing specific requirements for genomic tests such as paternity, ancestry, personalized medicine, and genetic compatibility that can be performed on fully sequenced genomes.[7] The authors carefully examined today's in vitro procedures, while analyzing their security and privacy requirements in the digital domain, succeeding in gradually designing privacy-preserving specialized protocols with reasonable overhead.

However, miscommunications between bioinformatics and privacy experts have also created several technical gaps, such as the use of de-identification in the futile attempt of guaranteeing genome donors' privacy, or the design of impractical cryptographic techniques for computation over encrypted data that aren't resilient to sequencing errors. On one hand, genomics and bioinformatics researchers lack the necessary skills to effectively formalize adversarial and threat models or to rigorously evaluate the security of proposed countermeasures. On the other hand, cryptography and security experts often make unrealistic assumptions about computational, business, or operational models and tend to oversimplify the practical requirements of real-world genomic applications.

Security and privacy practitioners have a unique opportunity in genomics to proactively contribute expertise and lessons learned from previous efforts to protect individuals' medical privacy to a new field where existing threats are broadly known but rapidly evolving. In the wake of Internet and mobile technology revolutions, security and privacy were only considered afterward, leaving almost no choice other than deploying "patches" on top of existing infrastructures. In contrast, with genomics, we might still be able to apply a privacy-by-design approach, whereby privacy and data protection compliance is effectively designed into systems holding private information right from the start, rather than being bolted on afterward or ignored.[8] ■

References

1. E. Ayday et al., "The Chills and Thrills of Whole Genome Sequencing," to appear in *Computer*, 2014; http://arxiv.org/pdf/1306.1264.
2. M. Humbert et al., "Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy," *Proc. ACM Conf. Computer and Communications Security*, 2013, pp. 1141–1152.
3. A.L. McGuire et al., "To Share or Not to Share: A Randomized Trial of Consent for Data Sharing in Genome Research," *Genetics in Medicine*, vol. 13, no. 11, 2011, pp. 948–955.
4. M. Gymrek et al., "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, no. 6117, 2013, pp. 321–324.
5. Y. Erlich and A. Narayanan, "Routes for Breaching and Protecting Genetic Privacy," arXiv preprint 1310.3197, 2013.
6. D. Mascalzoni et al., "Informed Consent in the Genomics Era," *PLoS Medicine*, vol. 5, no. 9, 2008, p. 192.
7. P. Baldi et al., "Countering GATTACA: Efficient and Secure Testing of Fully-Sequenced Human Genomes," *Proc. ACM Conf. Computer and Communication Security*, 2011, pp. 692–702.
8. A. Cavoukian, "Privacy by Design," Information and Privacy Commissioner of Ontario, Canada, 2009; www.privacybydesign.ca.

Emiliano De Cristofaro is a senior lecturer at University College London. His research interests include privacy, security, and applied cryptography. De Cristofaro has a PhD in networked systems from the University of California, Irvine. Contact him at me@emilianodc.com.