

# The Chills and Thrills of Whole Genome Sequencing

Erman Ayday<sup>1</sup>   Emiliano De Cristofaro<sup>2</sup>   Jean-Pierre Hubaux<sup>1</sup>   Gene Tsudik<sup>3</sup>

<sup>1</sup> EPFL, [firstname.lastname@epfl.ch](mailto:firstname.lastname@epfl.ch)

<sup>2</sup> PARC (a Xerox Company), [me@emilianodc.com](mailto:me@emilianodc.com)

<sup>3</sup> UC Irvine, [gts@ics.uci.edu](mailto:gts@ics.uci.edu)

## Abstract

In recent years, whole genome sequencing (WGS) evolved from a futuristic-sounding research project to an increasingly affordable technology for determining complete genome sequences of complex organisms, including humans. This prompts a wide range of revolutionary applications, as WGS is a promising means for improving modern healthcare and providing a better understanding of the human genome, in particular its relation to diseases and response to treatments. However, this progress raises worrisome privacy and ethical issues, since, besides uniquely identifying its owner, the genome contains a treasure trove of highly personal and sensitive information.

In this article, after summarizing recent advances in genomics, we discuss some important privacy issues associated with human genomic information and identify a number of particularly relevant research challenges.

## 1 Introduction

Recent years have witnessed impressive advances in DNA sequencing. Both throughput and affordability of new-generation sequencing platforms have increased at a pace faster than Moore's Law would otherwise predict. It seems quite reasonable to assume that, in a few years, most individuals in developed countries will have the means of having their genomes sequenced, thus enabling personalized genomic medicine and facilitating preventive treatment and diagnosis.

However, for now this remains only a prospect and much more research is needed to understand the very complex relationship between genome and health. To conduct this research, the scientific community needs

large cohorts of patients (or volunteers) willing to share their genetic material. One example is the Personal Genome Project, wherein all participants must agree to have their genomic data – along with other personal information – made publicly available on the Internet. This clearly raises many potential privacy, ethical, and legal concerns.

The first documented case of privacy issues dates back to the end of the 19th century. It was triggered by the availability of a new and revolutionary observation and identification tool: the photo camera. Since then, several other such tools have become widespread, including: video cameras, credit cards, Web browsers, and mobile phones. These tools reveal our presence and habits in various spheres of life, as well as our communication and mobility patterns. DNA sequencing greatly exacerbates this problem, as the genome represents our ultimate biological identity. By combining genomic data with information about one's environment or lifestyle (often easily obtainable from social networks), could make it possible to infer that individual's phenotype.

In general, access to genomic data prompts some important privacy concerns: (i) DNA reflects information about genetic conditions and predispositions to specific diseases such as Alzheimer's, cancer, or schizophrenia, (ii) DNA contains information about ancestors, siblings, and progeny, (iii) DNA (almost) does not change over time, hence revoking or replacing it (as with other forms of identification) is impossible, and (iv) DNA analysis is already being used both in law enforcement and healthcare, thus prompting numerous ethical issues. Furthermore, it is hard to assess or estimate the extent of the personal information that could be extracted or derived from the genome in the future. (At the same time, it does not take a great leap of faith to believe that it will be impressive).

In this article, after a brief summary of some basic genomic concepts, we describe some expected benefits of personalized medicine and discuss notable privacy issues, as well as associated research challenges.

## 2 Background

The human genome is encoded in double stranded DNA molecules consisting of two complementary polymer chains. Each chain consists of simple units called nucleotides (A,C,G,T). The DNA of a person can be retrieved from various sources (e.g., saliva, hair, skin, blood). Once a sample is collected, the genetic material is extracted and then sequenced – using a DNA sequencing platform – to obtain the so-called raw DNA sequence. This is usually in the form of short reads, each including hundreds of nucleotides from random parts of the genome. Next, the raw reads are quality-controlled, analyzed, and aligned to the reference genome (digital nucleic acid sequence database, assembled by scientists as a representative example of our species’ set of genes), allowing the progressive reconstruction of the whole sequenced genome. After further analysis of these short reads, eventually, the 3.2 billion letters in the DNA sequence of the person are reconstructed.

Even though most of the DNA sequence is conserved across the whole human population, around 0.5% of each person’s DNA (which corresponds to several millions of nucleotides) is different from the reference genome, owing to genetic variations. Single nucleotide polymorphism (SNP) is the most common DNA variation. A SNP is a position in the genome holding a nucleotide that varies between individuals. Currently, there are approximately 40 million confirmed SNPs in the human population [24] and this number is increasing very rapidly. Multiple Genome-Wide Association Studies (GWAS) performed in recent years have shown that a patient’s susceptibility to particular diseases can be partially predicted from sets of his SNPs [20, 10]. For example, it was reported that there are three genes bearing a total of ten particular SNPs necessary to (partially) analyze susceptibility to Alzheimer’s disease [29]. Thus, leakage of SNPs poses a significant threat to genomic privacy.

An interesting characteristic of the SNPs, called *Linkage Disequilibrium* (LD) [11], poses a notable privacy threat. LD is observed whenever SNPs are not independent of each other. Therefore, the content of

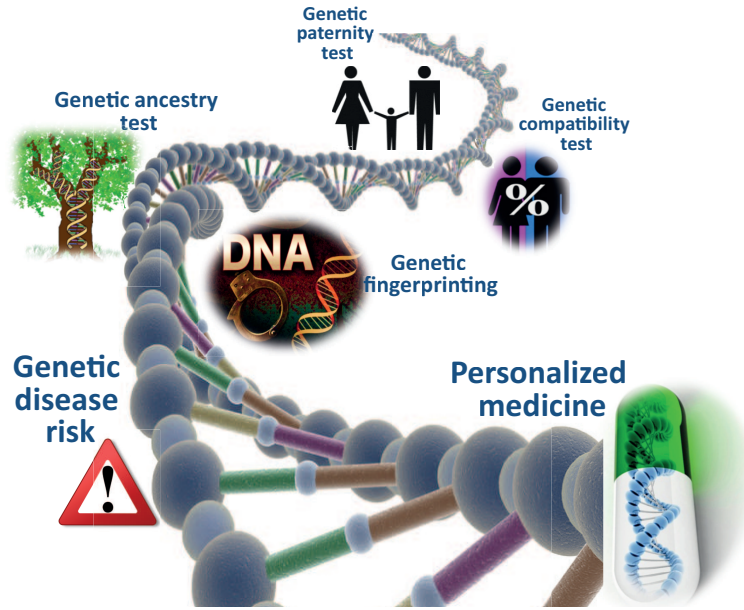
a SNP position (i.e., nucleotides residing at that SNP position) can be inferred from the contents of other SNP positions using the LD relationship. The most well-known example of the aforementioned threat is the ApoE status of Jim Watson (the co-discoverer of DNA), who published his genome with the exception of his ApoE gene (which carries SNPs to determine the risk for Alzheimer’s disease). However, it was later shown that these SNPs on his ApoE gene can be (probabilistically) inferred using their LD relationships with the published ones [26].

## 3 Towards Personalized Medicine, and Beyond

Widespread and affordable availability of fully sequenced human genomes creates enormous opportunities, which we summarize in Fig. 1 (and discuss in this section).

In particular, whole genome sequencing (WGS) facilitates the advent of a new era of predictive, preventive, participatory, and personalized medicine (“P4 medicine”) [19]. **Personalized genomic medicine** is recognized as a significant paradigm shift and a major trend in health care [34], where treatment and medication type/dosage would be tailored to the precise genetic makeup of individual patients.

For instance, certain genetic mutations are known to alter drug metabolism, thus genomic tests are often used today to predict a patient’s response to particular drugs. The study of the impact of genetic variations on the response to medications is called **pharmacogenomics**. A well-known example in this direction includes testing for SNP mutations in the *tpmt* gene for childhood leukemia patients, prior to prescribing 6-Mercaptopurine and Azathioprine drugs. The *tpmt* gene codes for the TPMT enzyme that metabolizes these drugs. Moreover, genetic polymorphisms affecting enzymatic activity of TPMT are correlated with variations in sensitivity and toxicity response to such drugs. Other common examples include pre-testing for Zelboraf (Roche’s treatment for advanced skin cancer), as well as pre-treatment testing for Philadelphia chromosome mutations related to Acute Lymphoblastic Leukemia (ALL) and BRCA1/BRCA2 genes in correlation to familial breast and ovarian cancer syndromes. Experts estimate that about a third of the 900 cancer drugs currently in clinical trials could soon



**Figure 1:** Applications of genomics.

come to market with a recommendation for a DNA or other molecular test attached [6].

Furthermore, Vanderbilt University’s PREDICT program (Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment) [23] helps physicians tell which drugs are most likely to work for their patients, and which they should avoid, based on the genetic characteristics of the patients, instead of long trial and error periods. For instance, [35] reports how a specific cholesterol-lowering drug was successfully selected based on the genomic profile of a patient with coronary artery disease.

Experts predict that advances in WGS will further stimulate the development of personalized medicine [15]. Commercial companies, like Knome, already offer services that take raw genome data and create usable reports for doctors. In general, the availability of a patient’s fully sequenced genome will enable clinicians, doctors, and testing facilities to run a number of complex, correlated genetic tests in a matter of seconds, using specialized computational algorithms (as opposed to more expensive and slower *in vitro* tests).

The democratization of low-cost whole genome genotyping and sequencing provides individuals with direct access to their genomic information, including to some **genetic disease risk tests**. For example, a well-known commercial company, 23andMe [1], provides relatively low-cost genetic disease risk tests for

960,000 specific SNPs. In Fig. 2, we illustrate the genetic disease risk results of a real human with fictional name “Greg Mendel” (whose genomic data is publicly available in the “demo mode”) provided by 23andMe. This table mainly shows the diseases for which Greg Mendel’s calculated risk is higher relative to average. In [32], Topol mentions a few real stories about how the disease risk values obtained from 23andMe helped early diagnosis of serious diseases (e.g., prostate cancer).

Name	Confidence	Your Risk	Avg. Risk
Atrial Fibrillation	★★★★	33.9%	27.2%
Prostate Cancer ♂	★★★★	29.3%	17.8%
Alzheimer’s Disease	★★★★	14.2%	7.2%
Age-related Macular Degeneration	★★★★	11.1%	6.5%
Colorectal Cancer	★★★★	7.8%	5.6%
Chronic Kidney Disease	★★★★	4.2%	3.4%
Restless Legs Syndrome	★★★★	2.5%	2.0%
Parkinson’s Disease	★★★★	2.2%	1.6%

**Figure 2:** Genetic risks of Greg Mendel for several diseases (source: 23andMe).

The availability of whole human genomes will also facilitate a number of genetic tests that today are performed *in vitro* by reducing costs and time. For instance, **genetic paternity tests** might be run very efficiently in computation, by designing algorithms that emulate *in vitro*, are highly accurate, and court ad-

missible. Furthermore, **ancestry and genealogical testing** is already offered by several commercial entities, which use publicly available genomic data from individuals belonging to different ethnic groups, and compare them against their customers' genomic information to understand how they relate to known ethnic groups. Similarly, **genetic compatibility tests**, which let potential or existing partners assess the risk of transmitting to their children a genetic disease with Mendelian inheritance [22], are offered by various online services.

## 4 Privacy and Ethical Pitfalls

While advances in whole genome sequencing are paving the way to extraordinary progress in healthcare (and beyond), they also prompt serious privacy, ethical, and security concerns. Besides uniquely identifying its owner, a genome contains information about one's ethnic heritage, predisposition to numerous physical and mental health conditions as well as other phenotypic traits [12, 7, 14]. We illustrate two main privacy threats to genomic data in Fig. 3.

Traditional approaches to privacy, such as de-identification or aggregation [21], are ineffective in the genomic context, since the genome itself is the ultimate identifier [18]. For instance, a recent study by Gymrek et al. [16] demonstrated the feasibility of re-identifying DNA donors from a public research database using information available from popular genealogy Web sites and other available information. Additional work on genomic re-identification includes [28] and [18].

The privacy problem is further exacerbated by the fact that genomes of any two closely related individuals are 99.9% identical, in contrast with 99.5%, on average, for two random people. Thus, the disclosure of one's genome leads to the leakage of significant genomic information about that person's close relatives, including parents, siblings and offspring. This is a problem regardless of how the disclosure occurs: voluntarily, accidentally or maliciously. This makes genomic privacy a unique problem since, in most other privacy-sensitive scenarios, only the individual's data is at stake, while in the genomic context, disclosure of personal information impacts a potentially large group of individuals. The most recent example of this issue is the controversy between the family members of deceased Henrietta Lacks (whose genome was se-

quenced and published after her death, without getting the permission of her family) and the scientists who are in favor of publishing genomes online for research purposes [31].

Even more worrying is that consequences of genomic data disclosure are not limited in time. In some other cases of leakage of one's private information, some recourse is possible. For example, bank account numbers and passwords can be changed, physical or electronic documents (even public key certificates) can be replaced and old ones can be revoked. In contrast, a genome is not mutable nor "revokable". Moreover, as large portions thereof are passed on to future generations, disclosure of one's genomic information can turn into an endless curse for both current and future generations.

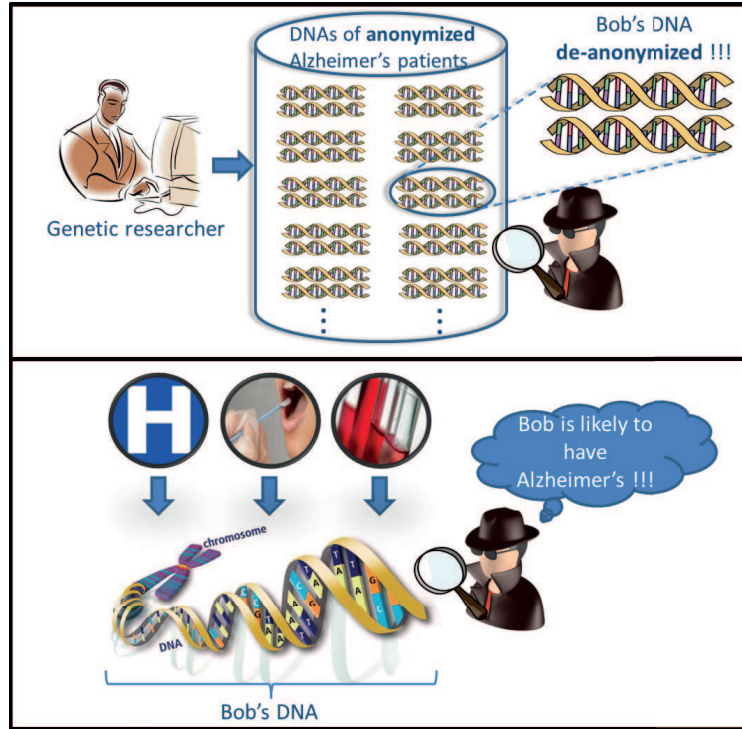
Based on the above, it is not surprising that privacy concerns represent a formidable obstacle for assembling large human genomic databases, e.g., for the purpose of conducting Genome-Wide Association Studies (GWAS). More generally, privacy concerns might actually stand in the way of advances in medicine and consequent improvements in overall healthcare. The same could apply in the domain of law enforcement where DNA-based identification is being increasingly used and there is a need for secure and reliable handling of large numbers of genomes.

On the part of the US federal government, there has been awareness of privacy and ethical issues in genomics. For example, as early as 1990, the National Human Genome Research Institute (NHGRI) established the Ethical, Legal and Social Implications (ELSI) Research Program with the goal of exploring repercussions of advances in genetic and genomic research on individuals, families and communities.

Federal laws, such as the 2003 Health Insurance Portability and Accountability Act (HIPAA), provide a general framework for protecting and sharing Protected Health Information (PHI). Furthermore, the Genetic Information Nondiscrimination Act (GINA) adopted in 1998 prohibits discrimination on the basis of genetic information with respect to health insurance and employment [33]. Also, some states, like California, have recently started to consider DNA privacy laws [30].

Even the popular culture, via sci-fi movies and literature, has touched upon genetic discrimination. For instance, the notion of **genism** that originated in the 1997 movie "GATTACA", denotes the theory that dis-





**Figure 3:** Main threats to human genomic data: (i) de-anonymization of the DNA donors from a public research database, and (ii) leakage of genomic data, and hence leakage of privacy-sensitive information.

tinctive human characteristics and abilities are determined by genes, resulting in discrimination as pernicious as racism [2].

While providing general guidelines, current legislation does not offer sufficient technical information about safe and secure ways of storing and processing digitized genomes. We believe that security and privacy issues for genomic data (in the context of both individual genomes and databases thereof) are timely, important and relatively poorly understood.

Privacy practitioners and consumer organizations are strongly advocating the **need for more restrictive legislation** as a result of gaps in current policies. A recent report from the Presidential Commission for the Study of Bioethical Issues [27] analyzed advances of whole genome sequencing, and highlighted growing privacy and security concerns. This report makes a few privacy and security recommendations, including, unfortunately, de-identification.

At the policy level, challenges include the need for informed consent to guard against surreptitious DNA testing. Authorities and companies should obtain written permission from citizens before collecting, analyzing, storing or sharing their genetic information, e.g.,

preventing collection of hair or saliva samples and using them for unauthorized sequencing.

On the other hand, some academics fear that restrictive (privacy-friendly) measures could seriously hinder genomic research. Scientists typically sequence DNA from large numbers of people in order to determine genes associated with particular diseases. The informed consent restriction would mean that large genomic datasets could not be re-used to study a different disease; researchers would either need to destroy the data after each study, or track down all previously enrolled study participants for each new authorization. Also, the similarity of related individuals' genome raises doubts as to whether or not relatives should also provide consent.

Finally, the collection and analysis of human genomes does not arise only in the contexts of research studies and improved healthcare. It also comes up in increasingly popular commercial (for-profit) applications, which are not well-regulated so far. An example is [genepartner.com](http://genepartner.com), which claims to do matchmaking based on unclear genetic features.

## 5 Open Research Problems

As discussed above, advances in genomics will soon result in large numbers of individuals having their genomes sequenced and obtaining digitized versions thereof. This poses a wide range of technical problems, which we explore below.

**Storage and Accessibility: Genome at Rest.** Due to its sensitivity and size (about 3.2 billion nucleotides), one key challenge is where and how a digitized genome should be stored. It is reasonable to assume that an individual who requests (and likely pays for) genome sequencing should own the result, as is already the case with any other personal medical results and information. This raises numerous issues, including:

- (1) Should the genome be stored on one's personal devices, e.g., a PC or a smartphone? If so, what, if any, special hardware security features (e.g., tamper-resistance) are needed?
- (2) Can it be outsourced to a cloud provider?
- (3) Should the sequencing facility keep an escrowed copy of the genome?
- (4) Should it be entrusted to one's personal physician and/or health insurance provider?
- (5) How is it to be stored: in the clear or encrypted? If the latter, where are encryption keys generated: at the lab? at owner's premises? at the cloud provider? Where are these keys stored?
- (6) How to guarantee integrity and authenticity of the digitized genome (i.e., guarantee that it corresponds to its actual owner)?
- (7) Should backups be made? If so, how often and where can copies be kept?
- (8) How can one erase a genome securely?
- (9) Should an individual periodically re-sequence their genome to take advantage of more accurate technology?

**Privacy: Genome in Action.** Given the genome's sensitivity, an individual should, ideally, never disclose any information contained therein. However, this would prevent the access to any genomic application

that cannot be entirely and securely performed *in situ*, i.e., within a secure perimeter of one's own personal device. In principle, this might be possible if operations are performed in some standardized and certified form. For example, if testing for a genetic disease requires matching a well-known pattern in some approximate location in the genome, that pattern and its parameters can be certified by some trusted agency (such as the US Food and Drug Administration). Thus, an individual could be assured that a legitimate test for a specific genetic disease is being conducted and the result is clearly communicated to that individual; the latter would then have the option to keep the result private.

Due to the sensitivity of genomic data, research on the privacy of genomic data has accelerated over the past few years. We can put this research in four main categories: (i) private string searching and comparison [5], (ii) private release of aggregate data [36], (iii) private alignment of raw genomic data [9], and (iv) private clinical usage of genomic data (e.g., for personalized medicine) [3, 4].

Nonetheless, it is hard to foresee the range and complexity of future genetic operations: some (future) tests might be too computationally complex to be performed within the confines of a personal device. Furthermore, some genetic testing would probably involve multiple genomes, e.g., when tracing origins of some conditions, siblings or parents/children might need to be tested together. Similarly, in assessing risks of genetic conditions for future progeny, both prospective parents have to be tested. Also, some genetic tests often constitute a trademark of a pharmaceutical/biomedical company, whose intellectual property needs to be protected – for more details, see [25, 17, 8].

As soon as genomic information leaves the (virtual) hands of its owner, purely technical approaches to privacy become insufficient. Legal and professional guidelines would certainly be needed to govern how information is transmitted, stored, processed, and eventually disposed of on the receiving end, e.g., by the physician, hospital, pharmacist or medical lab.

**Long-term data protection.** Even if genomes are encrypted (at rest), encryption schemes considered strong today might gradually weaken in the long term, whereas genome sensitivity does not dissipate over time. It is not too far-fetched to imagine that a third-party in possession of an encrypted genome might be able, e.g., 30 or 50 years later, to de-

crypt it. For instance, the Advanced Encryption Standard (AES) scheme supports key lengths up to 256 bits – a key length estimated by NIST, following Moore’s law, to be secure several years after 2030 (<http://www.keylength.com/en/4/>). However, computational breakthroughs or newly discovered weaknesses of the encryption algorithm might allow breaking the encryption earlier than expected. Also, even many years from now, the leakage of an individual’s genome, although deceased, may still severely affect the privacy of his or her progeny.

If one assumes that an encrypted genome will not be copied, then the genome could be periodically re-encrypted in time. Alternatively, one could split the genome, using secret-sharing techniques, and partition it among several providers; however, it remains an open question how to efficiently reassemble the genome for various operations as well as how to guarantee non-collusion between providers.

**Accuracy and Accountability:** Computational genomic tests should guarantee accuracy comparable to current analog *in vitro* equivalents. For example, a software implementation of the paternity test on digitized genomes should offer at least the same confidence as its *in vitro* counterpart currently admissible in a court of law. Also, computational tests should aim at accountability, e.g., by providing lasting guarantees that tests are run correctly and on intended genomic information.

**Efficiency.** Computational genomic tests should incur minimal communication/computational costs. Arguably, minimality in this setting is relative to the context of such tests. For instance, patients may be inclined (and used) to wait several days to obtain results of genetic tests that concern their health, however in the computational setting, long running times on personal devices might hinder the real-world practicality of these tests (besides taking out one of the main motivations for computational tests.)

**Usability.** Computational genomic tests that involve end-users should be usable by, and meaningful to, regular non-tech-savvy individuals. This translates into non-trivial questions, such as: how much understanding should be expected from a user running a test? What information (and at what level of granularity) should be presented to the user as part of a test and as its outcome? Do privacy perceptions and concerns experienced by patients correspond to what the scien-

tific community would expect? For instance, one may think that patients will be likely to trade off privacy of their genomes to enable tests that can save them from, e.g., cancer. However, only little work, e.g., [13] has focused on users’ concerns and (mis)proven common beliefs to this regard, thus pointing out the need for ethnographic studies in the field. Also, it remains an open problem to explore effective ways to communicate to the users the potential privacy risks associated with genomic information and its disclosure.

**Large-scale research on human genomes.** As discussed in Section 4, potential privacy, legal, and ethical concerns appear to conflict with enabling large-scale research on human genomes, such as Genome-Wide Association Studies (GWAS). Such scale is however required for researchers to discover associations between genetic make-up and medical conditions. One current trend is to store donors’ genomes in clouds and use analytics techniques running on powerful computer clusters. Once again, this prompts many privacy and legal concerns (also relevant to relatives and descendants).

## 6 Conclusion

This paper discussed some “chills and thrills” of an emerging phenomenon – affordable and readily available genomic sequencing. As something radically novel, it brings great opportunities and significant concerns, especially (as we have shown) pertaining to personal privacy. Mitigating privacy issues will require long-term collaboration among geneticists, other healthcare providers, ethicists, lawmakers, and computer scientists. We believe that this collaboration will not occur naturally and, in order to foster it, funding agencies need to target this topic. Until recently, at least in the United States, genomic privacy unfortunately fell into a sort of a “funding gap” between several agencies. One obvious candidate for playing a key funding role is the National Institute of Health (NIH). Yet, although it covers bioinformatics, NIH has not funded privacy research in the genomic context. The National Science Foundation (NSF), the main agency responsible for funding academic computer-science research, recently initiated a “Smart and Connected Health” program that includes so-called “integrative projects” requiring collaboration among computer and health sciences. It remains to be seen whether this

program will engender long-range genomic privacy research. Other US funding agencies have not, thus far, targeted genomic privacy. A similar situation can be observed in Europe: of course, there are numerous EU and nationally funded projects focusing on e-health, some of which address data protection. However, the genomic privacy challenge has been vastly overlooked, and the number of computer scientists working on the topic is even lower than in the United States. An additional issue is that, although most officials in charge of data protection typically have a strong legal background, they lack computer science expertise. As a consequence, they tend to rely on legislation more than on technology.

In conclusion, if not addressed, privacy issues highlighted in this article could affect numerous individuals. Moreover, by impeding genomic research, it would also affect the well-being of our society as a whole. Thus, we believe that there is an urgent need for collaboration among researchers in the fields outlined above.

## 7 Acknowledgements

EA and JPH would like to thank Dr. Jacques Fellay, Dr. Paul J. McLaren, Dr. Jacques Rougemont, and Dr. Amalio Telenti for their useful comments and suggestions. They would also like to express their gratitude to Dr. Vincent Mooser, Dr. Didier Trono and Dr. Martin Vetterli for their encouragements in this research endeavor.

## References

- [1] 23andMe. <https://www.23andme.com/>.
- [2] G. Annas. Genism, Racism, and the prospect of genetic genocide. In *World Conference Against Racism, Racial Discrimination, Xenophobia and Related Intolerance*, 2001.
- [3] E. Ayday, J. Raisaro, and J. Hubaux. Privacy-enhancing technologies for medical tests using genomic data. (*short paper*) in *20th Annual Network and Distributed System Security Symposium (NDSS)*, 2013.
- [4] E. Ayday, J. Raisaro, P. McLaren, J. Fellay, and J. Hubaux. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. *Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech)*, 2013.
- [5] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. In *CCS*, 2011.
- [6] A. Burke. Foundation Medicine: Personalizing Cancer Drugs. [http://is.gd/foundation\\_medicine](http://is.gd/foundation_medicine), 2012.
- [7] T. Canli. The emergence of genomic psychology. *Nature*, 8, 2007.
- [8] T. Caulfield, R. M. Cook-Deegan, F. S. Kieff, and J. P. Walsh. Evidence and anecdotes: an analysis of human gene patenting controversies. *Nature biotechnology*, 24(9), 2006.
- [9] Y. Chen, B. Peng, X. Wang, and H. Tang. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *NDSS*, 2012.
- [10] Eupedia. Genetically inherited traits, conditions, and diseases. [http://www.eupedia.com/genetics/medical\\_dna\\_test.shtml](http://www.eupedia.com/genetics/medical_dna_test.shtml).
- [11] D. S. Falconer and T. F. Mackay. *Introduction to Quantitative Genetics (4th Edition)*. Addison Wesley Longman, Harlow, Essex, UK, 1996.
- [12] J. Fowler, J. Settle, and N. Christakis. Correlated genotypes in friendship networks. *Proceedings of the National Academy of Sciences*, 108(5), 2011.
- [13] U. Francke, C. Dijamco, A. K. Kiefer, N. Eriksson, B. Moiseff, J. Y. Tung, and J. L. Mountain. Dealing with the unexpected: consumer responses to direct-access BRCA mutation testing. *PeerJ*, 1, 2013.
- [14] M. Fumagalli et al. Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. *Experimental Medicine*, 206(6), 2009.
- [15] G. Ginsburg and H. Willard. Genomic and personalized medicine: foundations and applications. *Translational Research*, 154(6), 2009.
- [16] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.
- [17] N. Hawkins. The Impact of Human Gene Patents on Genetic Testing in the UK. *Genet. Med.*, 13(4), 2011.
- [18] N. Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8), 2008.
- [19] L. Hood and D. Galas. P4 Medicine: Personalized, Predictive, Preventive, Participatory A Change of View that Changes Everything. [http://www.cra.org/ccc/docs/init/P4\\_Medicine.pdf](http://www.cra.org/ccc/docs/init/P4_Medicine.pdf), 2009.
- [20] A. D. Johnson and C. J. O'Donnell. An open access database of genome-wide association results. *BMC Medical Genetics* 10:6, 2009.
- [21] B. Malin. An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *Journal of the American Medical Informatics Association*, 12(1), 2005.
- [22] V. McKusick and S. Antonarakis. *Mendelian inheritance in man: a catalog of human genes and genetic disorders*. John Hopkins University Press, 1994.
- [23] My Drug Genome. Using Genetics to Personalize Medication Treatment. <http://www.mydruggenome.org/overview.php>.



- [24] National Center for Biotechnology Information. dbSNP. <http://www.ncbi.nlm.nih.gov/projects/SNP/>.
- [25] National Human Genome Research Institute. Intellectual Property and Genomics. <http://www.genome.gov/19016590>.
- [26] D. Nyholt, C. Yu, and P. Visscher. On Jim Watson’s APOE status: Genetic information is hard to hide. *European Journal of Human Genetics*, 17, 2009.
- [27] Presidential Commission for the Study of Bioethical Issues. PRIVACY and PROGRESS in Whole Genome Sequencing. <http://www.bioethics.gov/cms/sites/default/files/PrivacyProgress508.pdf>, 2012.
- [28] L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, and E. D. Green. The Complexities of Genomic Identifiability. *Science*, 339(6117):275–276, 2013.
- [29] S. Seshadri et al. Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA*, 303, 2010.
- [30] H. Shen. California considers DNA privacy law – Academic researchers fear measures would prohibit work with genetic databases. <http://www.nature.com/news/california-considers-dna-privacy-law-1.10677>, 2012.
- [31] R. Skloot. The Immortal Life of Henrietta Lacks, the Sequel. <http://nyti.ms/ZiMp2J>.
- [32] E. Topol. *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*. Basic Books, 2012.
- [33] M. Wadman. Genetics bill cruises through senate. *Nature*, 453, 2008.
- [34] A. Weston and L. Hood. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *Journal of Proteome Research*, 3(2), 2004.
- [35] K. Whitney. PREDICT helps pinpoint right statin for patient. <http://news.vanderbilt.edu/2012/10/predict-helps-pinpoint>.
- [36] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang. To release or not to release: Evaluating information leaks in aggregate human-genome data. In *ESORICS*, 2011.